HANDBOOK OF WEATHER, CLIMATE, AND WATER

ATMOSPHERIC CHEMISTRY, HYDROLOGY, AND SOCIETAL IMPACTS

THOMAS D. POTTER
BRADLEY R. COLMAN

HANDBOOK OF WEATHER, CLIMATE, AND WATER

Atmospheric Chemistry, Hydrology, and Societal Impacts

Edited by

THOMAS D. POTTER

BRADLEY R. COLMAN

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Potter, Thomas D. 1929–
Handbook of weather, climate, and water: chemistry, hydrology, and societal impacts
Thomas D. Potter & Bradley R. Colman.
p. cm.
ISBN 0-471-21489-2 (cloth : alk paper)
1. Atmospheric physics. 2. Atmospheric chemistry. 3. Hydrology. I. Colman, Bradley R. II. Title

QC861.3 .P67 2002
551.6–dc21

Printed in the United States of America.

10 9 8 7 6 5 4 3 2

DEDICATION AND ACKNOWLEDGMENTS

Many people have assisted in the production of this Handbook—the Contributing Editors, the Authors, our editors at Wiley, friends too numerous to mention, and our families who supported us during the long process of completing this work. Professor Peter Shaffer, University of Washington, is owed deep appreciation for his untiring generosity in sharing his experience and talent to solve many problems associated with this large project. They all deserve much credit for their contributions and we want to express our deep thanks to all of them.

Finally, we want to dedicate this work to Tom Lockhart, the Contributing Editor of the Measurements part of the Handbook. Tom passed away in early 2001 and we regret that he will not be able to see the results of his efforts and those of his colleagues in final form.

Tom Potter and Brad Colman

PREFACE

The *Handbook of Weather, Climate, and Water* provides an authoritative report at the start of the 21st Century on the state of scientific knowledge in these exciting and important earth sciences. Weather, climate, and water affect every person on earth every day in some way. These effects range from disasters like killer storms and floods, to large economic effects on energy or agriculture, to health effects such as asthma or heat stress, to daily weather changes that affect air travel, construction, fishing fleets, farmers, and mothers selecting the clothes their children will wear that day, to countless other subjects.

During the past two decades a series of environmental events involving weather, climate, and water around the globe have been highly publicized in the press: the Ozone Hole, Acid Rain, Global Climate Change, El Ninos, major floods in Bangladesh, droughts in the Sahara, and severe storms such as hurricane Andrew in Florida and the F5 tornado in Oklahoma. These events have generated much public interest and controversy regarding the appropriate public policies to deal with them. Such decisions depend critically upon scientific knowledge in the fields of weather, climate, and water.

One of two major purposes of the Handbook is to provide an up-to-date accounting of the sciences that underlie these important societal issues, so that both citizens and decision makers can understand the scientific foundation critical to the process of making informed decisions. To achieve this goal, we commissioned overview chapters on the eight major topics that comprise the Handbook: Atmospheric Dynamics, Climate System, Physical Meteorology, Weather Systems, Measurements, Atmospheric Chemistry, Hydrology, and Societal Impacts. Each of the sections was organized by a distinguished scientist who is a leading authority within that major field. In addition to writing an overview chapter, this scientist served as the Contributing Editor for that section of the Handbook. Each Contributing Editor selected both the topics and authors of the individual chapters, thus ensuring that the most important material has been included. The chapter authors are themselves leading experts in their specialty. These overview chapters present, in terms understandable to everyone, the basic scientific information needed to appreciate the major environmental issues listed above.

The second major purpose of the Handbook is to provide a comprehensive reference volume for scientists who are specialists in the atmospheric and hydrologic

areas. In addition, scientists from closely related disciplines and others who wish to get an authoritative scientific accounting of these fields should find this work to be of great value. The 95 professional-level chapters are the first comprehensive and integrated survey of these sciences in over 50 years, the last being completed in 1951 when the American Meteorological Society published the Compendium of Meteorology.

The *Handbook of Weather, Climate and Water* is organized into two volumes containing eight major sections that encompass the fundamentals and critical topic areas across the atmospheric and hydrologic sciences. This volume contains sections on the highly important topics of Atmospheric Chemistry, Hydrology, and Societal Impacts. The section on Atmospheric Chemistry contains thorough descriptions of the major biogeochemical cycles (carbon, oxygen, nitrogen and sulfur) that describe how chemical elements and compounds are transferred between the atmosphere, oceans, land and the biosphere, and their relationship to important environmental issues such as global climate change, the ozone hole, acid rain, and air pollution. The Hydrology section includes in-depth discussions of all parts of the hydrologic cycle (rain, snow, evaporation, runoff, ground water, and soil moisture), plus chapters on floods, remote sensing and GIS in hydrology, and stochastic processes in hydrology. Societal Impacts has chapters on the social effects of all of the major environmental issues.

To better protect against weather, climate, and water hazards, as well as to promote the positive benefits of utilizing more accurate information about these natural events, society needs improved predictions of them. To achieve this, scientists must have a better understanding of the entire atmospheric and hydrologic system. Major advances have been made during the past 50 years to better understand the complex sciences involved. These scientific advances, together with vastly improved technologies such as Doppler radar, new satellite capabilities, numerical methods and computing, have resulted in greatly improved prediction capabilities over the past decade. Major storms are rarely missed nowadays because of the capability of numerical weather-prediction models to more effectively use the data from satellites, radars and surface observations, and weather forecasters' improved understanding of threatening weather systems. Improvements in predictions are ongoing. The public can now rely on the accuracy of forecasts out to about five days, when only a decade or so ago forecasts were accurate to only a day or two. Similarly, large advances have been made in understanding the climate system during the past 20 years. Climate forecasts out to a year are now made routinely and users in many fields find economic advantages in these climate outlooks even with the current marginal accuracies, which no doubt will improve as advances in our understanding of the Climate System occur in future years.

Tom Potter and Brad Colman

Color images from this volume are available at ftp://ftp.wiley.com/public/sci_tech_med/weather/.

CONTRIBUTORS

RICHARD ARIMOTO, Carlsbad Environmental Monitoring and Research Center, New Mexico State University, 1400 University Drive, Carlsbad, NM 88220-3575

ROGER C. BALES, University of Arizona, Department of Hydrology and Water Resources, Tucson, AZ 85721-0011

ABDELLATIF BENCHERIFA, Techlink, Advanced Technology Park, 900 Technology Boulevard, Snite A, Bozeman, MT 59718-6857

MICHELE M. BETSILL, Colorado State University, Department of Political Science, Fort Collins, CO 80523-1782

KEITH BEVEN, Lancaster University, Department of Environmental Science, Lancaster LA1 4YQ, United Kingdom

J. D. BRADSHAW, Georgia Institute of Technology, Earth and Atmospheric Sciences, Atlanta, GA 30332

KENNETH BROAD, International Research Institute for Climate Prediction 61, Route 9W, Monell Building Palisades, NY 10964-8000

PAOLO BURLANDO, Institute of Hydromechanics and Water Resouces, ETH-Hönggerberg, Zurich, Switzerland

STANLEY A. CHANGNON, Illinois State Water and Changnon Climatol, 801 Buckthorn, Mahomet, IL 61853

G. CHEN, Georgia Tech, 22 Bobby Dodd Way NW, Room 205A, Atlanta, GA 30332

M. CHIN, National Aeronautics and Space Administration, Goddard Space Flight Center, Mail Code 916, Greenbelt, MD 20771

DON CLINE, National Weather Service, NOAA, National Operational Hydrologic Remote Sensing Center, Chanhassen, MN 55317-8582

STEWART J. COHEN, University of British Columbia, Sustainable Development Research Institute, 2029 West Mall, Vancouver, British Columbia, V6T 1Z2 Canada

J. H. CRAWFORD, NASA Langley Research Center, Mail Code 483, Hampton. VA 23681-0001

D. D. DAVIS, Georgia Institute of Technology, Earth and Atmospheric Sciences, Atlanta, GA 30332

THOMAS E. DOWNING, University of Oxford, Environmental Change Unit, Oxford, United Kingdom

MARY W. DOWNTON, National Center for Atmospheric Research, Box 3000, Boulder, CO 80301

EDWIN T. ENGMAN, NASA Goddard Space Flight Center, Laboratory for Hydrospheric Processes, Hydrological Sciences Branch, Code 974 Greenbelt, MD 20771

JACK FISHMAN, NASA Langley Research Center, Atmospheric Sciences Research, Hampton, VA 23681-2219

D. L. FREAD, National Weather Service, Office of Hydrology, 622 Stone Road, Westminster, MD 21158

MICHAEL H. GLANTZ, ESIG/NCAR, 3450 Mitchell Lane, Boulder, CO 80301

WILLIAM B GRANT, NASA Langley Research Center, Atmospheric Sciences Research, MS 401A, Hampton, VA 23681

WILLIAM L. GROSE, NASA Langley Research Center

EVE GRUNTFEST, University of Colorado, Department of Geography and Environmental Studies, 1420 Austin Bluffs Parkway, P.O. Box 7150, Colorado Springs, CO 80933-7150

HOSHIN GUPTA, University of Arizona, Department of Hydrology, Tucson, AZ 85721-0011

R. L HEATHCOTE, Finders University, Adelaide, Australia

PAUL R. HOUSER, NASA-GSFC, Hydrological Sciences Branch, Greenbelt, MD 20771

DANIEL J. JACOB, Harvard University, Department of Earth and Planetary Sciences, 29 Oxford Street, Cambridge, MA 02138

STEVEN JENNINGS, University of Colorado at Colorado Springs, 1420 Austin Bluffs Parkway, P.O. Box 7150, Colorado Springs, CO 80933-7150

JACK A. KAYE, NASA Langley Research Center, Office of Earth Sciences, Washington, DC 20546-0001

M. A. K. KHALIL, Portland State University, Department of Physics, P.O. Box 751 Portland, OR 97207-0751

WILLIAM P. KUSTAS, USDA Agricultural Research Service, Hydrology Laboratory, Beltsville, MD 20750

DONALD H. LENSCHOW, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000

MARCY E. LITVAK, University of California-Irvine, Earth System Science, Irvine, CA 92697-3100

S. C. LIU, Georgia Tech, Georgia Institute of Technology, Earth and Atmospheric Sciences, Atlanta, GA 30332

WILLIAM C. MALM, Colorado State University, Foothills Campus, Cooperative Institute for Research in the Atmosphere, Fort Collins, CO 80523-1375

NANDISH MATTIKALLI, Cambridge Research Associates, 1430 Spring Hill Road, Suite 200, McLean, VA 22102

PAULETTE MIDDLETON, Creator, President Panorama Pathways, http://PanoramaPathways.net; Rand Environment, Environmental Science and Policy Center, 2385 Panorama Avenue, Boulder, CO 80304

KATHLEEN A. MILLER, National Center for Atmospheric Research, Boulder, CO

MARIO J. MOLINA, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139-4307

M. SUSAN MORAN, Southwest Watershed Research Center, USDA Agricultural Research Service, 2000 East Allen Road, Tucson, AZ 85719

NEVILLE NICHOLLS, Bureau of Meteorology, GPO Box 1289K, Melbourne, Victoria 3001 Australia

JOHN M. NORMAN, University of Wisconsin, Department of Soil Science, 1525 Observatory Drive, Madison, WI 53706-1299

PAUL NOVELLI, NOAA Climate Monitoring and Diagnostics Laboratory, Environmental Research Laboratories, 325 Broadway, Boulder, CO 80303-3328

KENNETH E. PICKERING, University of Maryland, Department of Meteorology, 3433 Computer and Space Sciences Building, College Park, MD 20742-2425

ROGER A. PIELKE, JR., University of Colorado/CIRES, Campus Box 488, Boulder, CO 80309-0488

ROGER S. PULWARTY, U.S. Department of Commerce, NOAA Office of Global Programs, 1100 Wayne Avenue, Suite 1210 Silver Springs, MD 20910

JORGE A. RAMÍREZ, Colorado State University, Department of Civil Engineering, Fort Collins, CO 80523

JOSÉ D. SALAS, Colorado State University, Department of Civil Engineering, Fort Collins, CO 80523

STEPHEN H. SCHNEIDER, Stanford University, Department of Biological Sciences and Institute of Stanford, CA

STEPHEN E. SCHWARTZ, Brookhaven National Laboratory, ASD, Blgd. 815E, P.O. Box 5000, Upton, NY 11973-5000

ROGER A. SEDJO, Resources for the Future, 1616P Street NW, Washington, DC 20036

JOHN H. SEINFELD, California Institute of Technology, Pasadena, CA 91125

M. J. SHEARER, Portland State University, Department of Physics, P.O. Box 751, Portland, OR 97207-0751

SANFORD SILLMAN, University of Michigan, Atmospheric, Oceanic, and Space Sciences, 2455 Hayward, Ann Arbor, MI 48109-2143

JAMES A. SMITH, Princeton University, Department of Civil Engineering and Operations R, Princeton, NJ 08544

SOROOSH SOROOSHIAN, University of Arizona, Department of Hydrology and Water Resources, College of Engineering and Mines, Tucson, AZ 85721-0011

YOLANDE STOWELL, Environmental Resources Mangement, 8 Cavendish Square, London, W1M OER UK

WILL SWEARINGEN, 900 Technology Boulevard, Suite A, Bozeman, MT 59718-6857

ANNE M. THOMPSON, NASA-Goddard Space Flight Center, Laboratory for Atmospheres, Greenbelt, MD 20771

JUAN B. VALDÉS, University of Arizona, Department of Civil Engineering, Tucson, AZ 85721

COLEEN VOGEL, University of the Witwatersrand, School of Geography, Archaeology and Environment, Private Bag 3, Johannesburg, South Africa

CHRIS WALCEK, University of Albany, Atmospheric Sciences Research Center, 251 Fuller Road, Albany, NY 12203

M. L WESELY, Argonne National Laboratory, Environmental Research Division, 9700 South Cass Ave, 203 ER, Argonne, IL 60439

MARTHA P. L. WHITAKER, University of Arizona, Department of Hydrology and Water Resources, Tuscon, AZ 85271-0011

DONALD A. WILHITE, University of Nebraska, Lincoln, NB 68583-0749

WILLIAM W.-G. YEH, University of California, Department of Civil and Environmental Engineering, 5731/5732 Boelter Hall, Box 951593, Los Angeles, CA 90095-1593

IGOR S. ZONN, ESIG/NCAR, P.O. Box 3000, Boulder, CO 80307

CONTENTS

SECTION 2 HYDROLOGY

Contributing Editor: Soroosh Sorooshian

SECTION 1

ATMOSPHERIC CHEMISTRY

CHAPTER 1

OVERVIEW: ATMOSPHERIC CHEMISTRY

JACK FISHMAN

The study of atmospheric chemistry focuses on how chemical constituents cycle through the atmosphere. Excluding water vapor (which can account for as much as 2 to 3% of the volume of the atmosphere under extremely moist conditions), more than 99.9% of the remaining dry atmosphere is comprised of nitrogen (78.1%), oxygen, (20.9%), and argon (0.93%). Unlike the study of conventional meteorology, where the atmosphere is generally treated as a bulk medium, atmospheric chemistry focuses on each individual constituent (commonly referred to as trace gases) and the chemical reactions that take place among them.

When discussing atmospheric chemistry, it is perhaps most convenient to separate the discussion into two distinct chemical regimes: the stratosphere and the troposphere. In the stratosphere, the most important trace gas is ozone, O_3, whereas in the troposphere, it can be argued that one of the most important trace gases is carbon dioxide, CO_2. Both of these trace gases are intimately tied to the issue of global change as measurements over the past several decades confirm that stratospheric ozone is decreasing and that carbon dioxide is increasing. Ozone in the stratosphere is vital for shielding the biosphere from harmful ultraviolet radiation; a decrease in the amount of ozone in the stratosphere will result in damage to biota at the ground. On the other hand, carbon dioxide is an important trace gas (second in importance to water vapor) that keeps infrared radiation within the lower atmosphere, and it is generally agreed that an increase in CO_2 may have important climatic implications and lead to global warming.

The source of energy that drives the chemical processes in the atmosphere is the same source that drives Earth's weather engine, namely the sun. Furthermore, the high-energy ultraviolet radiation emitted by the sun initiates a series of reactions in

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

the upper atmosphere as these high-energy photons break the stable molecules, N_2 and O_2, apart into their atomic components. This high energy not only is capable of breaking these very strong molecular bonds apart, but it is also capable of stripping away electrons creating a source of ions in the atmosphere above ~ 50 km. This region of the atmosphere is called the ionosphere, and its chemistry will not be discussed in this section. For more information about the chemistry of the ionosphere, mesosphere, and thermosphere, see Brasseur and Solomon's (1986) *Aeronomy of the Middle Atmosphere*, Chapter 6 and various sections in Chapter 5. These ions and atoms can feed some of the chemical cycles that take place in the stratosphere, such as supplying reactive nitrogen species (e.g., see Fig. 1).

From an atmospheric chemistry point of view, important cycles take place in both the stratosphere and the troposphere; this section will concentrate on the chemistry taking place in these regions of the atmosphere. To a certain extent, the chemistry of the stratosphere is somewhat less complex than the chemistry in the troposphere because only large-scale meteorological processes are present at these high altitudes; smaller scale processes such as precipitation can be generally neglected. Also important is the fact that the sources of trace species in the stratosphere are not determined from small-scale sources and can thus can be quantified using a simplified methodology.

In the stratosphere, observing and gaining an understanding of how the distribution of ozone evolved was the primary research emphasis from the 1930s through the 1960s. Understanding how its abundance and distribution has been perturbed by anthropogenic inputs has been the focus of intense research efforts since the 1970s.

1 STRATOSPHERIC CHEMISTRY: UNDERSTANDING THE OZONE LAYER

Ozone was discovered in 1839 by the German scientist Christian Frederich Schönbein at the University of Basel in Switzerland. Because of its pungent odor, its name was taken from the Greek word *ozein*, meaning "odor." Schönbein's research, subsequent to his discovery, focused on verifying his hypothesis that ozone was a natural trace constituent of the atmosphere. As a result of interest in the late nineteenth century, there are a surprisingly large number of ambient measurements during that time.

The primary study of ozone focused on the chemistry of the stratosphere when it was hypothesized and then verified that most of Earth's ozone was located at an altitude of 20 to 50 km (also called the ozonosphere) high above Earth's surface. The British physicist Sir Sidney Chapman put forth the premise that sufficiently intense ultraviolet radiation [at wavelengths (λ); $\lambda < 242$ nm) breaks apart molecular oxygen into two oxygen atoms. This reaction is commonly written:

$$O_2 + h\nu \rightarrow O + O \qquad \lambda < 242 \text{ nm} \tag{1}$$

where $h\nu$ is the standard notation for a photon.

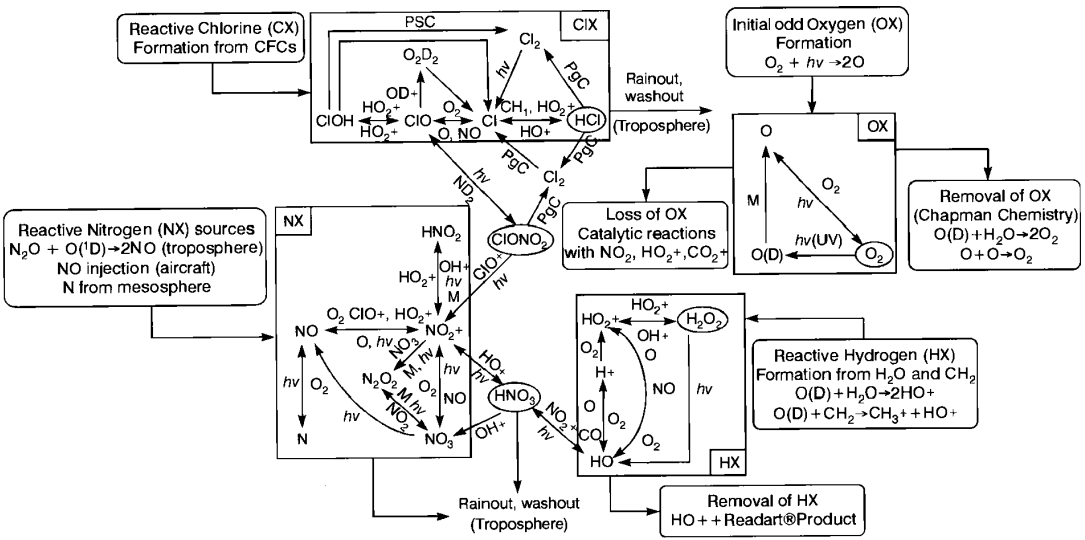


Figure 1 Simplified schematic diagram showing the interaction of the chemical families (large boxes) in the stratosphere. The round-cornered boxes define the sources and sinks for the chemical families. Reservoir (longer-lived) trace species are enclosed by ovals. Nitric acid (HNO_3) is a reservoir species for both reactive hydrogen (HX) and reactive (NX) families; chlorine nitrate ($ClONO_2$) is a reservoir species for the reactive chlorine (ClX) and NX families. The chemistry has been simplified by omitting bromine and iodine chemistry from the figure.

As the air becomes denser at lower altitudes in the stratosphere, most of this high-energy radiation is absorbed, and the oxygen molecules can no longer be broken apart. At these altitudes, the oxygen atoms will efficiently combine with the oxygen molecules and the formation of ozone occurs through the reaction:

$$O + O_2 + M \rightarrow O_3 + M \tag{2}$$

where M is a nonreactive third body that absorbs any excess collisional energy that may be present. Thus, there is a preferred region in the atmosphere where sufficient ultraviolet energy is concurrently present with the proper amount of molecular density to create ozone, and the altitude region at which these processes are most prevalent is commonly referred to as the ozone layer.

Ozone can also be photolyzed in the atmosphere by weaker ultraviolet radiation ($\lambda < 320$ nm) to give back molecular and atomic oxygen:

$$O_3 + h\nu \rightarrow O_2 + O(^1D) \qquad \lambda < 320 \text{ nm} \tag{3}$$

and also by visible radiation ($\lambda < 600$ nm) to yield atomic oxygen in its ground state, $O(^3P)$, rather than the more energetic $O(^1D)$ state; furthermore ozone can react with atomic oxygen (in either its ground or excited state) to give two molecules of oxygen:

$$O_3 + O \rightarrow 2O_2 \tag{4}$$

To complete the possible reactions in a "pure oxygen" atmosphere, two atoms of oxygen can combine in a three-body reaction to give molecular oxygen back to the system:

$$O + O + M \rightarrow O_2 + M \tag{5}$$

The set of five reactions involving only the various states of oxygen in the stratosphere are commonly referred to as "Chapman chemistry" and did a remarkable job of describing qualitatively why the ozone layer existed where it did. The speeds at which the five reactions took place in the atmosphere were measured independently in the laboratory and are called reaction rate constants (denoted k_4 for reaction 4, k_5 for reaction 5, etc.). Reaction rate constants are often temperature and pressure dependent. The rates of photolysis are noted by the letter j (e.g., j_3 for photolytic reaction 3, etc.) and are primarily dependent on the cross section of the individual molecule as a function of wavelength (those that have weaker bonds and can be broken apart more easily have larger cross sections) and the number of incident photons at those wavelengths (commonly called the photon flux).

As the field of chemistry progressed, other reactions were measured in the laboratory that were also believed to occur in the atmosphere with sufficient speed that they were eventually hypothesized to take an active role in the destruction and formation of ozone. These reactions dealt with derivatives of various forms of hydrogen in the

stratosphere. The chemistry of the stratosphere was modified accordingly to account for this new "wet photochemistry," which involved reactions being measured in the laboratory, was in the 1950s and 1960s. The rationale behind this new chemistry was that atomic oxygen, $O(^1D)$, could react with water vapor to form the hydroxyl radical, OH:

$$H_2O + O(^1D) \rightarrow 2OH \tag{6}$$

Another important source of reactive hydrogen in the stratosphere is degradation of methane CH_4, by $O(^1D)$. Regardless of the initial source of the OH radical, it could then react with ozone to form another radical, HO_2, the hydroperoxy radical, which can lead to a catalytic cycle that becomes an efficient mechanism by which ozone can be removed from the atmosphere:

$$OH + O_3 \rightarrow HO_2 + O_2 \tag{7}$$

$$\underline{HO_2 + O_3 \rightarrow OH + 2O_2} \tag{8}$$

$$2O_3 \rightarrow 3O_2 \quad \text{(net cycle)}$$

The above reactions helped to explain some of the observed differences between the measurements that were routinely made in the 1950s and 1960s and the calculated distribution of ozone determined from an oxygen-only atmosphere.

The next major modification to atmospheric chemistry came about from the inclusion of nitrogen chemistry into the reaction scheme of the stratosphere. Nitrous oxide, N_2O, was known to be a natural trace gas in the troposphere which did not have any identifiable removable mechanisms in lower atmosphere. Consequently, it could drift to the stratosphere where it was eventually attacked by the $O(^1D)$ atom to form nitric oxide, NO:

$$N_2O + O(^1D) \rightarrow 2NO \tag{9}$$

With the presence of NO in the stratosphere, another catalytic cycle of ozone destruction could occur through the following reaction sequence:

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{10}$$

$$\text{followed by} \quad \underline{NO_2 + O \rightarrow NO + O_2} \tag{11}$$

$$\text{Net cycle} \quad O_3 + O \rightarrow 2O_2$$

The importance of reactive nitrogen chemistry in the stratosphere was independently brought to light circa 1970 by Paul Crutzen, a recent Ph.D. in meteorology at the time from the University of Stockholm, and Harold Johnston, a chemistry professor at the University of California.

These catalytic ozone destruction cycles involving nitrogen and hydrogen species were the impetus behind the Climatic Impact Assessment Program (CIAP) of the

1970s, which became the rationale for determining the potential damage to the ozone layer that might result from flying a fleet of supersonic transport (SST) planes in the lower stratosphere. These planes would emit NO and H_2O directly into the stratosphere, and a confederation of U.S. federal agencies was charged with the task of determining how the ozone layer would be harmed by such a fleet. Although economic considerations eventually lay behind the decision for the United States not to pursue the development of a commercial fleet of SSTs, the environmental debate that developed during the early 1970s also contributed to the decision not to pursue the building of this new type of airplane.

But the environmental concern became even more of a reason to spend an increasing amount of money on stratospheric chemistry when Ralph Cicerone and Richard Stolarski, both at the University of Michigan in the early 1970s, introduced the possibility that chlorine chemistry might also provide another important means by which stratospheric ozone might be destroyed:

$$\mathrm{Cl} + \mathrm{O_3} \rightarrow \mathrm{ClO} + \mathrm{O_2} \tag{12}$$

$$\text{followed by} \quad \mathrm{ClO} + \mathrm{O} \rightarrow \mathrm{Cl} + \mathrm{O_2} \tag{13}$$

$$\text{Net cycle:} \quad \mathrm{O_3} + \mathrm{O} \rightarrow 2\mathrm{O_2}$$

Shortly after the chlorine cycle was identified as a potential mechanism for stratospheric ozone destruction, Mario Molina and F. Sherwood Rowland, both chemists at the University of California at Irvine, proposed that a group of anthropogenic chlorine-containing compounds could provide the source of significant amounts of chlorine in the stratosphere (Molina and Rowland, 1974). These compounds, known as chlorofluoro-carbons ($CFCl_3$ and CF_2Cl_2) were used primarily in air-conditioning systems and as propellants for aerosol spray cans that proliferated the use of these compounds in the 1960s. These substances had no known removal mechanism in the troposphere, and Molina and Rowland hypothesized that their only eventual sink would be drifting to the upper stratosphere where they would be destroyed by high-energy ultraviolet radiation resulting in the release of their reactive chlorine atoms into the chemistry of the stratosphere. Figure 1 shows the chemical reactions within each reactive family [e.g., the reactive nitrogen family (NX) the reactive hydrogen family (HX), etc.] and also how each of these individual chemical cycles would influence stratospheric ozone chemistry. The circled trace gas in each box in Figure 1 is the longest-lived species for that particular reactive group. Chlorine nitrate ($ClONO_2$) and nitric acid (HNO_3) are long-lived trace gases that serve as reservoirs of more than one reactive family.

As predicted, the buildup in chlorine led to a "thinning" of the ozone layer. Not predicted by the atmospheric chemists, however, was that the depletion of ozone intensified in the Antarctic stratosphere because of the unique meteorological conditions there. Stratospheric dynamics are such that an enhanced circulation develops during austral winter, which severely inhibits meridional heat exchange (unlike the Northern Hemisphere, where the position of major mountain ranges closer to the pole results in a more favorable situation for heat from middle and low latitudes to be

transported poleward). Thus, temperatures in the Antarctic lower stratosphere reach temperatures that are cold enough to allow for the formation of *polar stratospheric clouds* (PSCs) that provide ice surfaces that greatly perturb stratospheric chemistry by turning the long-lived (and relatively nonreactive) chlorine-containing compounds (chlorine nitrate, $ClONO_2$, and hydrochloric acid, HCl) into chlorine atoms, thereby greatly enhancing the destructive power of the reactive chlorine. Some of the main reactions that are influenced by PSCs are also shown in Figure 1 within the ClX box in the upper left of the figure. The net result has been the formation of the *ozone hole* whereby more than two-thirds of the normal amount of stratospheric ozone can be destroyed within a period few weeks as the austral winter ends (see Chapter 21, "Stratospheric Ozone Observations"). This phenomenon was first identified from ozonesonde measurements made by Joe Farman of the British Antarctic Survey in the early 1980s (Farman et al. 1985). By the early 1990s, more than 80% of the chlorine in the atmosphere was determined to be of anthropogenic origin (see Figure 2).

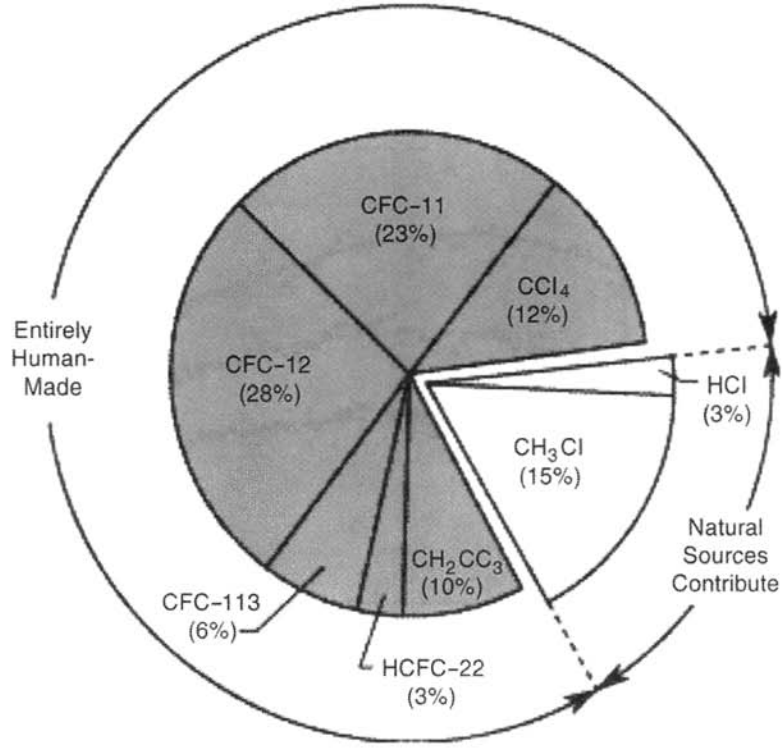


Figure 2 Diagram showing chlorine-containing compounds that release reactive chlorine to the stratosphere and the percentage that each contributes to stratospheric-reactive chlorine family (1994 estimates). Compounds that are completely produced by humans are shaded. Approximately 82% of the chlorine present in the stratosphere is of anthropogenic origin.

The environmental problem of stratospheric ozone depletion was successfully addressed by an international treaty in 1987 referred to as the Montreal Protocol, whereby a plan was set forth to phase out and eventually eliminate the manufacture and use of primary ozone-depleting chlorinated compounds (see Albritton et al., 1999). Figure 3 shows how the amount of man-made chlorine has decreased as a result of the effort to minimize the destruction of the ozone layer. As the amount of chlorine goes down in the stratosphere, model predictions suggest that the ozone hole should return to its pre-1980s level by the second or third decade of the new millennium. As a result of their important work on understanding the ozone layer and the chemical processes that drive the formation and destruction of ozone in both the stratosphere and the troposphere, Paul Crutzen, Mario Molina, and F. Sherwood Rowland were awarded the Nobel Prize for chemistry in 1995, the first time that atmospheric chemists received this coveted award. Whereas Rowland and Molina were both trained as chemists, Crutzen is the first meteorologist to receive the Nobel Prize.

Despite the complexity of Figure 1, it has been simplified by excluding the chemistry of two other halogen compounds: bromine and iodine. Reactive family chemistry of these halogens is similar to that shown for the reactive chlorine family.

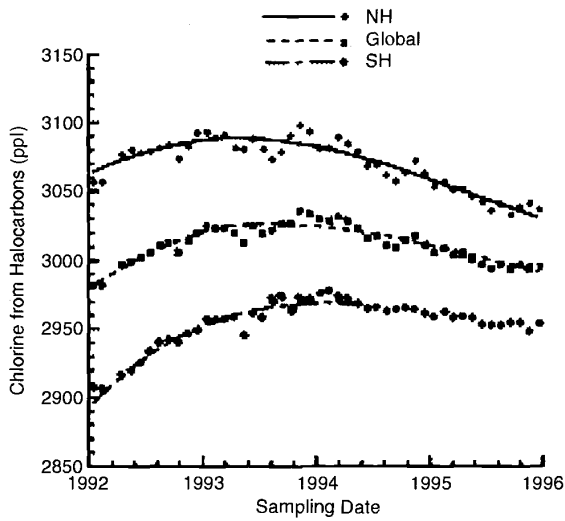


Figure 3 Monthly hemispheric and global tropospheric chlorine content measured between 1992 and 1996. Because of the international effort to curb human-produced chlorinated gases outlined in the Montreal Protocol of 1987 and subsequent amendments to that original agreement, the amount of chlorine in the troposphere showed an annual global decrease for the first time in the early 1990s. It is expected that stratospheric chlorine content will decline in the early 2000s, at which point the amount of ozone in the stratosphere should end its long-term declining trend.

Anthropogenic bromine compounds (called halons) are used as fumigants in agriculture and as a fire retardant on clothing. The most abundant iodine compound is methyl iodide, CH_3I, and has been observed in the atmosphere as a biomass burning product.

2 TROPOSPHERIC CHEMISTRY: A COMPLEX INTERACTION OF BIOGEOCHEMICAL CYCLES

Somewhat analogous to the chemical cycles just described for the stratosphere, atmospheric chemistry in the troposphere can also be viewed as a complex interaction of chemical cycles. These cycles, however, involve direct interaction with the biosphere and are also greatly complicated by the presence of the more complicated meteorology found only in the lower atmosphere. The biosphere serves as the exclusive source of carbon, and the carbon cycle, in turn, has important linkages that transcend land, ocean, and air.

The dominant form of carbon in the atmosphere is its completely oxidized state, carbon dioxide, which is present in the atmosphere at concentrations of ~ 360 ppmv (parts per million, by volume). Between the time carbon is stored in the biosphere and eventually becomes CO_2, many interesting chemical transformations and interactions take place. The second most abundant carbon-containing trace gas is methane, with an atmospheric concentration of ~ 1.8 ppmv, the only other atmospheric trace gas that exists in the concentrations of more than 1 ppmv, even in regions far removed from its sources. From an atmospheric chemistry point of view, methane plays an important role because its oxidation is closely linked to other carbon-containing trace compounds such as carbon monoxide (CO) and formaldehyde (CH_2O). The role of carbon dioxide, on the other hand, is not directly tied to chemical reactions taking place in the atmosphere, but rather to assessing how natural and anthropogenic processes contribute to increasing the global carbon burden.

The carbon cycle that couples the atmosphere with the biosphere is one of several important biogeochemical cycles in the Earth system. These cycles describe how specific elements and compounds are transferred between the principal global reservoirs—the atmosphere, land, oceans, and biosphere. Chemical and physical processes and transformations determine the partitioning of a material among the reservoirs. For example, for a fixed amount of carbon in these reservoirs, a certain fraction is found in the atmosphere, another fraction in the oceans, and so on; these fractions depend on the way the carbon is transferred between the reservoirs, the sizes of the reservoirs, and other factors. The amount to be found in the atmosphere, where it may have the most substantial effect on climate, is thus determined by the carbon's overall biogeochemical cycle.

The size of these cycles is enormous. For example, the carbon cycle transfers more than 10^{15} g of carbon per year from the atmosphere to other reservoirs. The terrestrial biosphere is dominated by trees and other vegetation; the amount of land biomass in animal form is much smaller, by approximately a factor of 100. Living

biomass on land is also ~100 times as massive as the total living biomass in the oceans. Dead biomass is even more plentiful than living biomass. Inanimate organic matter accumulates as plant litter in forests, building up as a layer of humic soil or peat. Eventually, the organic debris is consumed by microorganisms as it decays, or it may be burned; in either case, carbon dioxide is put into the atmosphere. The details of these cycles are complex, and many of the chapters in this section focus on specific key trace gases that are parts of these cycles. Other chapters focus on specific processes that are responsible for the conversion of trace gases to other species that may or may not remain in gaseous form. These processes, however, may provide the dominant vehicle through which various elements are transferred between the various reservoirs within the biogeochemical cycle (e.g., sulfur being removed from the atmosphere and transferred to the land through the formation of acid rain).

In addition to carbon dioxide and methane, the important members of the carbon cycle are carbon monoxide (CO), and a host of nonmethane hydrocarbons [also called volatile organic compounds (VOCs)] consisting of more than one carbon atom and a number of hydrogen atoms (C_nH_m, where n and m are integers). Other important cycles in Earth's atmosphere include nitrogen, oxygen, and sulfur. In the nitrogen (N) cycle, the key compounds are nitrogen (N_2), nitrous oxide (N_2O), nitrogen oxides (NO_x), consisting primarily of nitric oxide (NO) and nitrogen dioxide (NO_2), nitric acid (HNO_3), and the nitrate ion $(NO_3)^-$. Both molecular nitrogen and nitrous oxide are key elements of the biogeochemical nitrogen cycle, although neither is involved in any chemical reactions in the troposphere.

The oxygen cycle in the troposphere is comprised nearly exclusively of molecular oxygen (O_2), and ozone (O_3); note that atomic oxygen is not an important player in the troposphere despite its importance on the stratosphere. The primary players in the sulfur cycle are sulfur dioxide (SO_2), carbonyl sulfide (COS), hydrogen sulfide (H_2S), dimethyl sulfide $(CH_3)_2S$, sulfuric acid (H_2SO_4), and the sulfate ion $(SO_4)^{-2}$.

This section on atmospheric chemistry will include separate articles on reactive nitrogen species, tropospheric ozone, and atmospheric sulfur, as well as the carbon compounds carbon monoxide and methane and nonmethane hydrocarbons. A broad discussion will now be presented on the carbon, nitrogen, and oxygen cycles so that the reader will be better able to envision how the individual trace gases and processes fit into the "big picture."

3 GLOBAL CARBON CYCLE

Measurements of the concentration of air trapped in Antarctic ice cores indicate that over the last 200,000 years, atmospheric concentrations of CO_2 have fluctuated between 200 and 280 ppmv, until the last century. Data for the period AD 1000 and 1800 indicate that the concentration was quite stable, averaging 280 ppmv and varying over that period by only about 10 ppmv, indicating that the CO_2 cycle was in equilibrium in the centuries prior to the industrial revolution. Over the past 200 years, however, the concentration has increased from 280 to more than

360 ppmv; this increase is attributed primarily to burning of fossil fuels (primarily at northern middle latitudes) and tropical biomass burning. Through an examination of the isotopic composition of the CO_2 in the ice core samples, it can likewise be shown that nearly all of this increase is a result of fossil fuel combustion. Carbon has two isotopes, one with a molecular weight of 12 and the other with a molecular weight of 13. Naturally occurring carbon dioxide that has been put into the atmosphere through photosynthesis will be comprised of ~1% of the heavier ^{13}C; CO_2 that has been put into the atmosphere from fossil fuel burning will be slightly depleted in ^{13}C. Analysis of the isotopic $^{13}CO_2$-to-$^{12}CO_2$ ratio from air trapped in ice cores indicates that a smaller fraction of the CO_2 released to the atmosphere before the industrial revolution came from fossil fuel sources when compared to the modern-day ratios.

Currently, approximately 6 GtC (1 GtC = 10^{12} kg of carbon) are released to the atmosphere as a result of fossil fuel combustion. Between about 1850 and the early 1970s, the release of CO_2 increased exponentially at a relatively constant rate of 4.3% per year (Fig. 4). Since the oil crisis of 1973, the concerted efforts to reduce energy consumption have slowed the trend, but, nonetheless, an upward trend still continues, especially since the late 1980s. The cumulative production of CO_2 from fossil fuel is estimated to be 225 GtC, or ~30% of the current amount of CO_2 in the atmosphere. The result of this increase to the atmosphere is reflected in both ambient

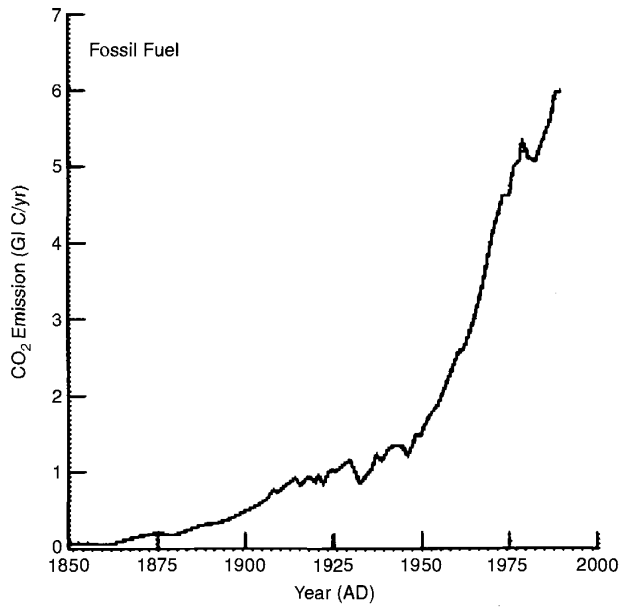
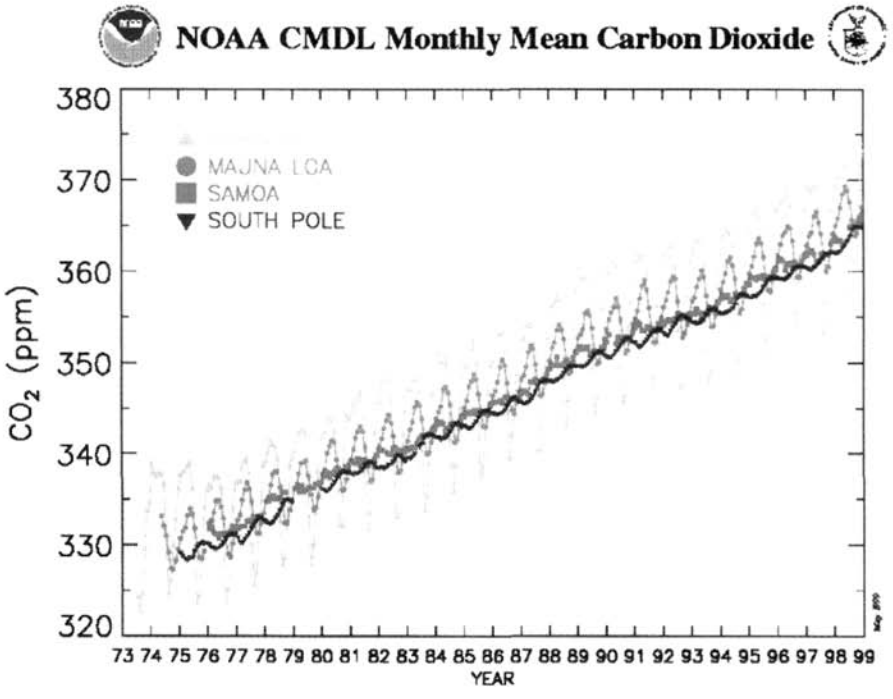


Figure 4 Estimates of CO_2 released to the atmosphere from fossil fuel combustion from 1850 to the present.



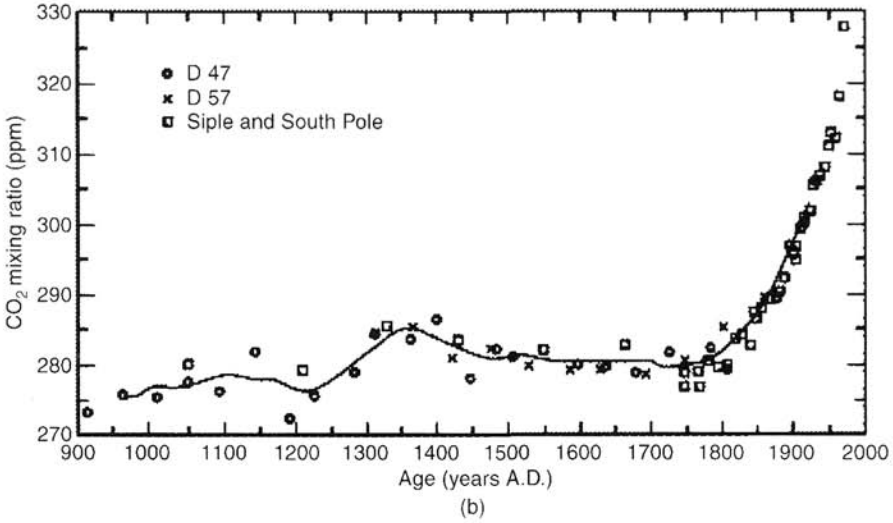


Figure 5 (see color insert) (*a*) Monthly concentrations of CO_2 measured from gas samples at four monitoring sites operated by NOAA's Climate Monitoring and Diagnostics Laboratory from the early 1970s; (*b*) CO_2 concentrations determined from ice core samples estimated to go back ~1000 years. See ftp site for color image.

measurements from several monitoring sites of NOAA's Climate Monitoring and Diagnostics Laboratory (see Fig. 5a) and from ice core data (Fig. 5b). The current rate of CO_2 increase is ~1.8 ppmv per year.

4 GLOBAL CARBON BUDGET

Balancing the global carbon budget is a challenging effort, but large strides have been made in recent years. One of the largest problems is the difficulty of measuring carbon fluxes over large scales as well as accurately modeling atmospheric and oceanic transport of carbon species. The global carbon cycle is shown schematically in Figure 6. In this figure, the numbers in the reservoirs are given in GtC and the fluxes in GtC per year. This figure shows that the *gross* exchange flux between the ocean and the atmosphere (on the order of 90 GtC/yr), and between the terrestrial biosphere and the atmosphere (on the order of 60 GtC/yr) are at least an order of magnitude larger than the CO_2 emissions from fossil fuel burning (5.5 GtC/yr) and from deforestation (net of 1.6 GtC/yr). On the other hand, the total anthropogenic input of CO_2 to the atmosphere (7.1 GtC/yr) is significant compared to the *net* exchange fluxes between the three carbon reservoirs. In particular, the net CO_2 flux into the terrestrial biosphere is the subject of an ongoing debate. Instead of attempting to deal with the complete carbon budget, it is easier to limit this discus-

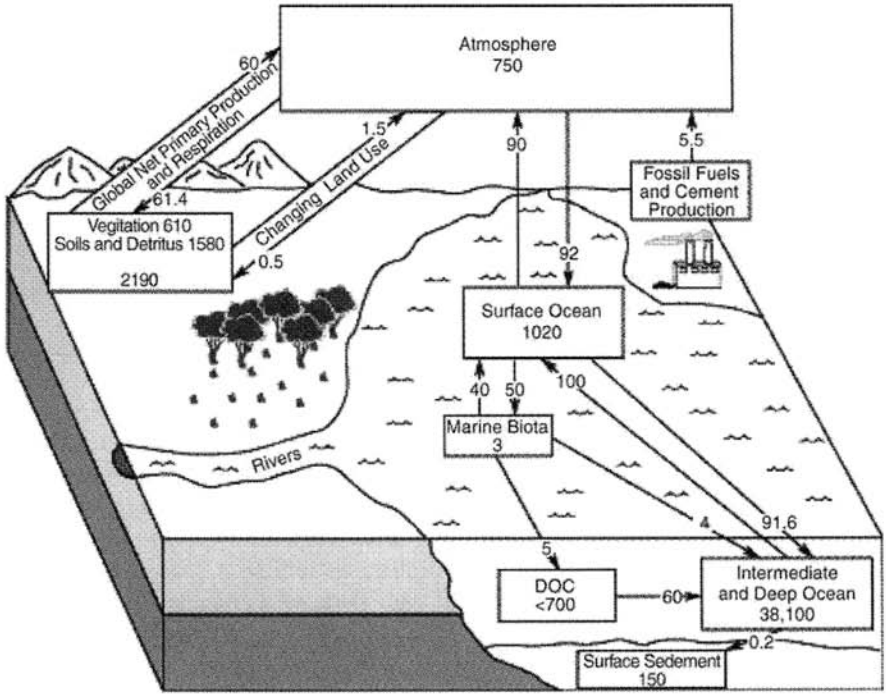


Figure 6 Schematic diagram showing the size of the carbon reservoirs and the amount of annual exchange between them.

sion to the *perturbation budget* for CO_2 (i.e., what happens to the CO_2 injected into the atmosphere by the combination of fossil fuel and biomass combustion and land-use change).

The single value in the global cycle that is known most accurately is the change of atmospheric CO_2 concentration, which has been measured continuously and averaged over a number of sites; it corresponds to 3.3 ± 0.2 GtC/yr. Fossil fuel combustion can also be estimated fairly accurately (through a knowledge of global coal and petroleum production) at a value of 5.5 ± 0.5 GtC/yr. Estimates of CO_2 contributions resulting from deforestation, primarily in tropics, are quite uncertain, ranging from 0.6 to 2.5 GtC/yr. Using an average value of 1.6 ± 1.0 GtC/yr, the Intergovernmental Panel on Climate Change estimates the average sources of anthropogenic CO_2 to the atmosphere as 7.1 ± 1.1 GtC/yr. Since 3.2 ± 0.2 GtC/yr accumulate in the atmosphere, the remaining 3.9 GtC/yr must be reabsorbed either by the oceans or by the terrestrial biosphere. Current models calculate an oceanic uptake of 2.0 ± 0.8, leaving an imbalance (or "missing sink") of 1.4 ± 1.5 GtC/yr. A considerable amount of research in recent years has been directed toward partitioning this missing sink into ocean and land components.

Research on many of the individual scientific issues is currently being conducted by numerous scientists throughout the world and the scientists involved in this research transcend a number of disciplines such as atmospheric science, ecology, microbiology, and others. These scientists are brought together to foster research related to the issues involving global change through the International Geosphere–Biosphere Program (IGBP), under the sponsorship of the International Council of Scientific Unions. IGBP has several "core" programs and one of them is the International Global Atmospheric Chemistry (IGAC), which focuses on tropospheric chemistry. Much of the discussion in the remainder of this chapter and in the other chapters in this section discuss results and research that are linked to the goals of IGBP.

5 ATMOSPHERIC CHEMISTRY WITHIN GLOBAL CARBON CYCLE

Despite its dominance as a member of the global carbon cycle, carbon dioxide does not play any significant role in atmospheric chemistry. The most abundant carbon species that is an active player in atmospheric chemistry is methane (CH_4), which reacts in the troposphere with the hydroxyl radical (OH) to initiate a series of reactions that were shown to play important roles in the global cycles of a number of atmospheric trace species. The seminal work in tropospheric chemistry was published by Hiram Levy in 1971. Levy showed that the hydroxyl radical should exist in sufficient quantities in the troposphere to initiate a sequence of photochemical reactions that both produce and destroy a number of important tropospheric trace gases.

The initial formation of OH comes from the photolysis products of ozone in the troposphere:

$$O_3 + h\nu \rightarrow O_2 + O(^1D) \qquad \lambda < 320 \text{ nm} \tag{3}$$

followed by its reaction with water vapor:

$$H_2O + O(^1D) \rightarrow 2OH \tag{6}$$

Hydroxyl then reacts with methane to form a host of products:

$$CH_4 + OH \rightarrow CH_3 + H_2O \tag{14}$$
$$CH_3 + O_2 + M \rightarrow CH_3O_2 + M \tag{15}$$
$$CH_3O_2 + NO \rightarrow CH_3O + NO_2 \tag{16}$$
$$CH_3O + O_2 \rightarrow CH_2O + HO_2 \tag{17}$$
$$CH_2O + h\nu \rightarrow CO + H_2 \qquad \lambda < 360 \text{ nm} \tag{18}$$

This sequence, commonly referred to as methane oxidation was hypothesized to be a major source of formaldehyde (CH_2O) and carbon monoxide (CO). In addition, other radicals such as CH_3O_2 (methyl peroxy) and HO_2 (hydroperoxy) were formed and became important factors in the tropospheric ozone budget. Once formaldehyde (CH_2O) was formed, it could photolyze in an alternate pathway to produce even more reactive radicals:

$$CH_2O + h\nu \rightarrow HCO + H \qquad \lambda < 360 \text{ nm} \tag{19}$$

followed by

$$H + O_2 + M \rightarrow HO_2 + M \tag{20}$$
$$HCO + O_2 + M \rightarrow HO_2 + CO + M \tag{21}$$

Note that either photolysis sequence of CH_2O results in the formation of carbon monoxide.

Like CO_2, methane also absorbs infrared radiation and contributes to global warming. Ice core data show that atmospheric CH_4 concentrations remained relatively constant at about half of its present form for thousands of years before beginning to increase about 200 years ago from ~0.65 ppmv to ~1.8 ppmv (Fig. 7). Methane's atmospheric lifetime is ~8 years and its dominant removal mechanism is oxidation by OH.

Methane is produced in oxygen-deficient environments of Earth's surface (swamps, lakes, rice paddies, tundra, boreal marshes, etc.). Methane production in soils and oceans is the end product of a variety of reductive pathways during the decomposition of organic matter. Methane is also released by cattle, termites, and perhaps other insects, whereas coal mining, natural gas losses, and solid-waste burning are important anthropogenic sources. Large amounts of methane are also produced by biomass burning. The global methane budget is given in Table 1.

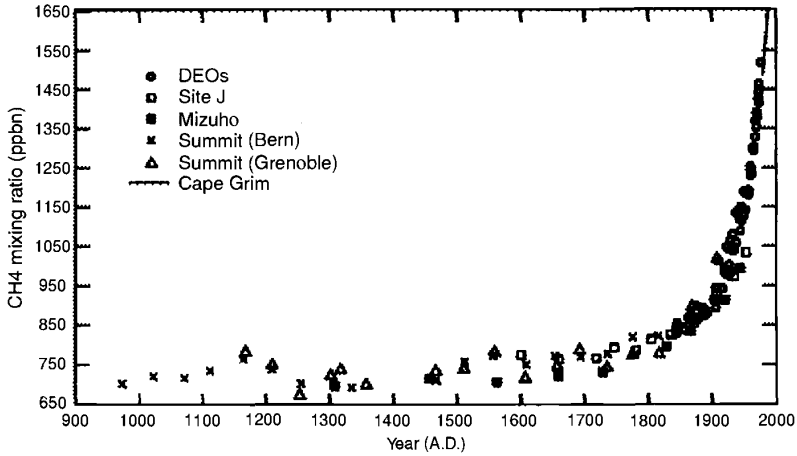


Figure 7 Ice for samples of atmospheric methane showing only a slight increase from 950 to 1800, but a sharp rise in concentration over the last 150 years.

One of the important products of methane oxidation is CO, which can also be oxidized by OH to form CO_2:

$$CO + OH \rightarrow CO_2 + H \tag{22}$$

Thus, CH_4, CO, and CO_2 are linked together through a series of oxidation processes that take place in the atmosphere; all forms of carbon emitted to the atmosphere eventually become CO_2.

As can be seen from the above discussion, carbon monoxide is caught in the middle as an intermediate oxidation product. One of the fundamental questions in tropospheric chemistry is to determine how much CO is emitted directly to the atmosphere, relative to the amount that is produced in situ through CH_4 oxidation. The primary and only significant sink for CO is removal by OH, which leads to an atmospheric residence on the order of 1 to 2 months. Thus, CO can be used as a useful tracer for atmospheric transport processes that take place on times scales of several days to a week or so.

6 CARBON MONOXIDE, NITROGEN OXIDES, AND OXIDIZING CAPACITY OF TROPOSPHERE

In some of the chapters that follow, there are detailed discussions on many individual trace gases such as CO, the oxides of nitrogen, and tropospheric ozone. Additionally, these three trace gases are of particular interest as they interact to a major degree to determine the oxidizing capacity of the troposphere. One of the most important

TABLE 1 Estimated Sources and Sinks of Methane

Sources	Magnitude (Tg CH_4 per year)	Range (Tg CH_4 per year)
Natural		
Wetlands	115	100–200
Termites	20	10–50
Ocean	10	5–20
Freshwater	5	1–25
CH_4 hydrate	5	0–5
Anthropogenic		
Coal mining, natural gas and petroleum industry	100	70–120
Rice paddies	60	20–150
Enteric fermentation	80	65–100
Animal wastes	25	20–30
Domestic sewage treatment	25	?
Landfills	30	20–70
Biomass burning	40	20–80
Total sources	515	
Sinks		
Atmospheric removal	470	450–520
Removal by soils	30	15–45
Atmospheric increase	32	28–37
Total sinks	532	

Source: Watson et al., 1992.

cycles that comes into play in the troposphere is the formation of ozone from carbon monoxide oxidation:

$$\mathrm{CO} + \mathrm{OH} \rightarrow \mathrm{CO_2} + \mathrm{H} \tag{22}$$

$$\mathrm{H} + \mathrm{O_2} + \mathrm{M} \rightarrow \mathrm{HO_2} + \mathrm{M} \tag{20}$$

$$\mathrm{HO_2} + \mathrm{NO} \rightarrow \mathrm{OH} + \mathrm{NO_2} \tag{23}$$

$$\mathrm{NO_2} + h\nu \rightarrow \mathrm{NO} + \mathrm{O} \qquad \lambda < 420 \text{ nm} \tag{24}$$

$$\mathrm{O} + \mathrm{O_2} + \mathrm{M} \rightarrow \mathrm{O_3} + \mathrm{M} \tag{25}$$

$$\mathrm{CO} + 2\mathrm{O_2} + h\nu \rightarrow \mathrm{CO_2} + \mathrm{O_3} \quad \text{(net cycle)}$$

This relatively simple catalytic cycle shows how the global budgets of CO and ozone in the troposphere are intertwined if there is a sufficient amount of NO and NO_2 present in the atmosphere. Since it has already been demonstrated that the CO

and CH_4 budgets are linked, one of the driving questions in atmospheric chemistry is the determination of what percentage of all these trace gases is natural, what fraction is anthropogenic, how have these budgets been perturbed over the past decades and centuries, and, lastly, how much will these budgets change in the coming decades. Through the complex interactions of these trace gases, the oxidizing capacity of the troposphere can be determined and possibly even predicted.

Earth is the only planet in this solar system where there is an oxidizing atmosphere, and although nearly all carbon emitted to the atmosphere eventually ends up as completely oxidized CO_2, and all hydrogen-containing trace species end up as H_2O, some interesting and often complex chemistry takes place along these oxidation pathways. Sulfur and nitrogen also eventually become oxidized, and the final products result in the formation of acids that, being soluble, contribute to the formation of acid rain. In recent years, chemical analysis of rain, fog, clouds, and dew have shown that the aqueous chemistry is as equally challenging and even more complicated than gas-phase chemistry. A complete understanding of aqueous-phase chemistry is still evolving, but many of the basic principles have become fairly well established and are described in more detail in some of the chapters in this section.

The most important species in clouds and precipitation is the hydrogen ion, whose concentrations can be indicated by specifying solution acidity, or pH. The presence of atmospheric CO_2 assures that nearly all atmospheric water droplets will be acidic; natural and anthropogenic nitrogen and sulfur increase the acidity (i.e., lower the pH value) to at least pH 5.0. Many urban areas experience pH levels nearer 4.0. Cloud and fog droplets are nearly always more acidic than rain, apparently because smaller cloud drop sizes inhibit dilution of the acidic constituents. In some fogs, the pH of the droplets has been measured as low as 1.7.

Organic compounds in the atmosphere also contribute to cloud acidity. Formaldehyde (CH_2O), often found in high concentrations where urban pollution is present, is a key tropospheric species with sufficiently high solubility that it can affect the acidity of rain. In remote regions, forests are known to emit large quantities of isoprene (C_5H_8), which can react with OH or O_3 to form more complex aldehydes (RCHO), and also resulting in the measurement of acid rain in the range of pH 5.0. Other carbon-, nitrogen-, and sulfur-containing acids also exist and are discussed in the chapters on acid rain, reactive nitrogen species, and sulfur species. The microphysics by which these trace gases are converted to both gaseous and aqueous forms of these acids is also treated explicitly in several chapters in this section.

Returning to the central theme of this overview, a considerable amount of research has been conducted over the past several decades to determine the budgets of carbon monoxide, methane, nitrogen oxides, and tropospheric ozone and to determine how these budgets are affected by human activity. Because these species do interact with each other, a series of different conclusions has been reached regarding human influence on tropospheric chemistry cycles. In the early 1970s, a series of research studies extrapolated Levy's initial hypothesis to postulate that many important chemical processes in the unpolluted atmosphere were dominated by (natural) methane chemistry. Since those initial studies, however, the atmospheric chemistry community has come to recognize that the natural background atmosphere and the concentrations of naturally occurring trace species such as methane, tropo-

spheric ozone, and carbon monoxide have increased dramatically, at rates comparable to, and, in most cases, even more than, carbon dioxide, the hallmark of proof of the concept of global change.

7 ATMOSPHERIC CHEMISTRY AND GLOBAL WARMING

The links of trace gas chemistry to climate change extend beyond the observed increase in CO_2 concentrations. Increases in CH_4, tropospheric O_3, and anthropogenically produced concentrations of the chlorofluorocarbons (CFCs) also contribute to global warming (see Fig. 8). If only the direct radiative effects of the trace gas increases are considered, 62% of the increase would be due to CO_2, 20% to CH_4, 4% to N_2O, and 14% to the CFCs. If chemical feedbacks are considered, the global warming due to changes with respect to changes in ozone concentrations (in both the stratosphere and troposphere) result in changes in ozone being as important as the changes in methane. With the international cooperation now in place to phase out the production and use of CFCs, future scenarios indicate that tropospheric ozone increases will replace methane as the second most important trace gas that contributes to the greenhouse effect (see Houghton et al., 1990; Houghton et al., 1996).

Furthermore, because tropospheric ozone is a relatively short-lived gas, especially when compared to other gases that contribute to global warming, increases in its concentration may have regional and seasonal effects that must be accounted for properly when temperature perturbations are being computed. Another consideration for regional climate change is the presence of particulates that are produced by fossil fuel combustion and biomass burning. Particles screen some of the incoming solar radiation and therefore result in a regional cooling effect. Studies to date show that both tropospheric ozone and man-made aerosols must be included in any scenarios attempting to simulate global climate and that the regional effects caused by these two constituents are comparable to the global perturbation of the longer-lived trace gases.

Lastly, the future buildup of methane and hydrogenated CFCs, the replacement gases to the CFCs, which can be removed from the atmosphere by OH oxidation, is complicated by the potential change in the oxidizing capacity of the troposphere. As O_3 increases in the troposphere, it is likely that the global abundance of OH will also increase, since the primary formation mechanism of OH in the troposphere is initiated by the photolysis of O_3 and the subsequent reaction of the excited oxygen atom with water vapor. If more OH is present, the removal of CH_4 and the replacements for the CFCs, which are removed in the troposphere by OH, becomes more efficient and thus their rate of increase is slowed.

8 STRATOSPHERE–TROPOSPHERE CHEMICAL AND CLIMATE INTERACTION

Although it is convenient to describe the chemistry of the stratosphere and the chemistry of the troposphere separately, both chemical and meteorological processes in one domain play an important role on the chemistry in the other. As previously

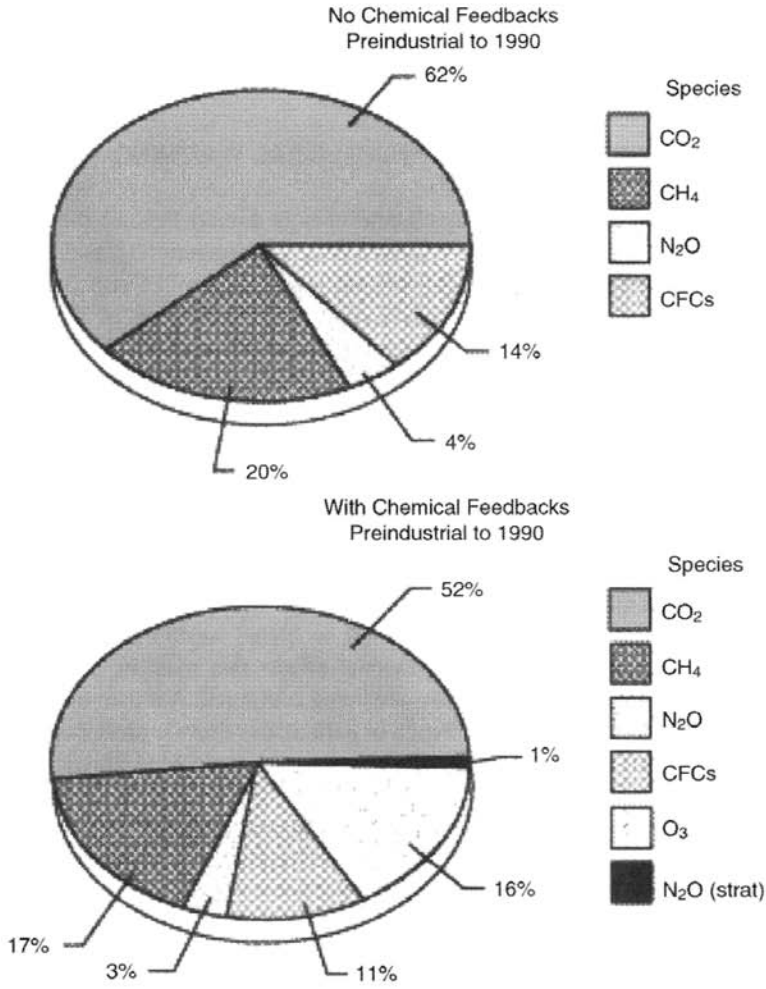


Figure 8 Contributions of various trace gases to global warming. Top graph shows the calculations using a model that does not include contributions from tropospheric ozone increases or from feedback relating to photochemical processes; bottom chart shows contributions using a model that includes photochemical feedback mechanisms. See ftp site for color image.

mentioned, one important aspect of tropospheric chemistry that impacts stratospheric chemistry is the removal of ozone-depleting chlorine in the troposphere before reaching the stratosphere.

Ozone depletion potentials (ODPs) provide a relative measure of the expected impact on ozone per unit mass emission of a gas compared to that expected from the same mass emission of CFC-11 integrated over time. Their primary purpose is for comparison of relative impacts of different gases upon ozone (e.g., for evaluating the relative effects of choices among CFC substitutes upon ozone). The two factors that

contribute to how effectively an anthropogenic compound depletes ozone is how much of it reaches the stratosphere and how much chlorine each molecule contains. CFC-11 (trichlorofluoromethane) contains three chlorine atoms and is not removed in the troposphere through chemical reactions. The primary replacements for many of these compounds contain a hydrogen atom, which can be attacked by OH in the troposphere. The lifetimes of these compounds are also presented in Table 2, which was compiled as part of the International Assessment of Ozone Depletion in 1994 (Albritton et al., 1995). Some of the compounds, such as HFC-134a, the primary choice for refrigerant in many automobile air-conditioning systems, contain no chlorine, and thus, virtually no chance of contributing to ozone depletion.

In addition to ODPs being established, the 1994 ozone assessment also produced a series of global warming potentials (GWPs) to provide a simple representation of the relative radiative forcing resulting from a unit mass emission of a greenhouse gas compared to a reference compound. Because of its central role in concerns about climate change, carbon dioxide has generally been used as the reference gas. The values presented in Table 2 are calculations based on a time horizon of 20 years. Evaluations of GWPs must also take into consideration the radiative property of the atmosphere at some future point, including the concentration of CO_2, and other major climate altering compounds such as nitrous oxide and methane. In the 1994

TABLE 2 Estimated Lifetime, Ozone Depletion Potential (ODP), and Global Warming Potential (GWP) for Various Anthropogenic Trace Gases

Trace Gas	Chemical Formula	Lifetime (years)	ODP	GWP
CFC-11	$CFCl_3$	50	1.0	5000
CFC-12	CF_2Cl_2	102	0.82	7900
CFC-113	$C_2F_3Cl_3$	85	0.90	5000
CFC-114	$C_2F_4Cl_2$	300	0.85	6900
CFC-115	C_2F_5Cl	1700	0.40	6200
Carbon tetrachloride	CCl_4	42	1.20	2000
Methyl chloroform	CH_3CCl_3	5.4	0.12	360
HCFC-22	CF_2HCl	13.3	0.04	4300
HCFC-123	$C_2F_3HCl_2$	1.4	0.014	300
HCFC-124	C_2F_4HCl	5.9	0.03	1500
HCFC-141b	$C_2F_3H_3Cl$	9.4	0.10	1800
HCFC-142b	$C_2F_3H_3Cl$	19.5	0.05	4200
HCFC-225ca	$C_3F_5HCl_2$	2.5	0.02	550
HCFC-225cb	$C_3F_5HCl_2$	6.6	0.02	1700
HCFC-134a	CH_2FCF_3	14	$< 1.5 \times 10^{-5}$	3300
HCFC-23	CHF_3	250	$< 4 \times 10^{-4}$	9200
HCFC-125	C_3HF_5	36	$< 3 \times 10^{-5}$	4800
Methyl bromide	CH_3Br	1.3	0.64	6200
Halon-1301	CF_3Br	65	12	
Halon-1211	CF_2HBr	20	5.1	

ozone assessment and the Intergovernmental Panel on Climate Change (IPCC) (Houghton et al., 1996) report, there are also GWPs calculated with time horizons of 100 and 500 years, and such calculations include even more uncertainty than the values presented here because of the assumed scenarios for emissions so far into the future. Thus, although some of the replacement compounds for the CFCs have a negligible impact on the ozone layer, they will make important contributions to the overall greenhouse effect caused by the emission of anthropogenic chemicals released to the atmosphere.

9 STRATOSPHERE–TROPOSPHERE EXCHANGE

The exchange of mass between the stratosphere and troposphere is important to the chemistry of both regions as it brings chemical species with sources in the troposphere (such as CFCs) into the stratosphere, while species with stratospheric origin (such as ozone) can be brought into the troposphere. Thus, the transport can be important for driving the chemistry in both regions. Analogous to the boundary layer being isolated from the free troposphere because of the presence of a substantial inversion, the troposphere is isolated from the stratosphere by the high static stability of the stratosphere. Similarly, just as the boundary layer is turbulent and well mixed compared to the free troposphere, the troposphere is relatively well mixed vertically and horizontally compared to the stratosphere. The mixing time in the troposphere (on the order of months within each hemisphere; on the order of a year between the hemispheres) is much shorter than the time required to exchange the mass of the entire troposphere with the stratosphere (on the order of 18 years, although due to the difference in mass the entire stratosphere mixes with the troposphere every 2 years).

The simplest way to visualize a model for stratosphere–troposphere exchange is to consider bulk exchange between the two domains accomplished by uniform rising motion across the tropical tropopause, poleward drift in the stratosphere, and by continuity of mass, a return flow into the troposphere at middle and high latitudes. Such a circulation was first proposed in the 1940s by Brewer to explain the observed low water vapor mixing ratios in the stratosphere. The only place near the tropopause where the temperature is low enough to accompany such low values of relative humidity is in the tropics, where the tropopause is high and cold. Dobson pointed out that poleward and downward advection of this type of mean circulation was consistent with the observed high concentration of ozone in the lower polar stratosphere, far from the region of photochemical production. Although the Brewer–Dobson model does not provide a complete description of the exchange process, it is believed to be essentially correct; see Holton et al. (1995).

The Brewer–Dobson circulation cell is now known to be predominantly wave driven. The morphology of stratospheric wave forcing indicates that upward movement of air into the stratosphere occurs in the tropics and downward movement of air into the troposphere occurs preferentially in winter in middle and high latitudes. Net cooling is required to transport air from the stratosphere into the troposphere

whereas net diabatic heating is required to transport air from the troposphere into the stratosphere. Extensive measurements during the STEP (Stratosphere–Troposphere Exchange Project) in the 1980s showed that specific vigorous convective events were primarily responsible for transporting tropical air into the stratosphere. Tropospheric air can be either mixed directly into the stratosphere when the cumulo-nimbus towers overshoot, mixed across the tropopause by turbulent motion, or moved upward due to radiative heating of cloud tops. The dehydration occurs because some or all of the condensed ice particles are returned to the troposphere by sedimentation while the dry air remains in the stratosphere. Soluble chemical species will be found in the ice particles rather than in the dry air surrounding them, so there may be a greater resistance to cross-tropopause transport of soluble compounds. A schematic diagram illustrating the general concept of the circulation between the troposphere and stratosphere is shown in Figure 9.

Mass flow from the stratosphere to the troposphere tends to be concentrated in dynamical events known as tropopause folds, in which the tropopause on the poleward side of the jet stream is distorted during the development of large-scale weather

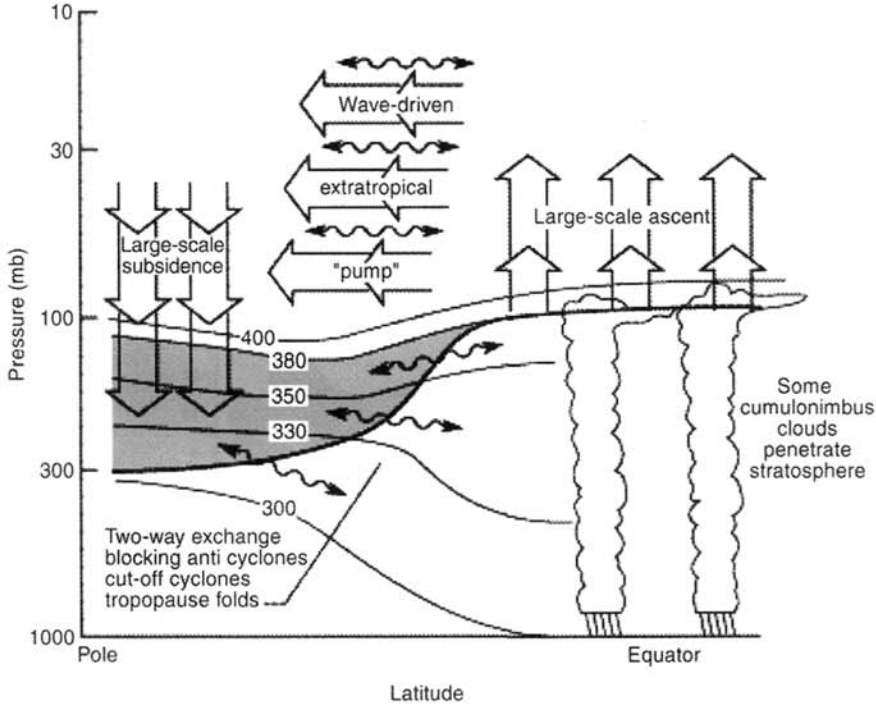


Figure 9 Schematic diagram showing the large-scale dynamical aspects of stratosphere–troposphere exchange. The wiggly double-headed arrows denote meridional transport by large-scale eddy processes. The broad arrows show transport by the global-scale circulation, which is the primary exchange mechanism that moves air across isentropic surfaces. (Reprinted with permission from Holton et al., 1995.)

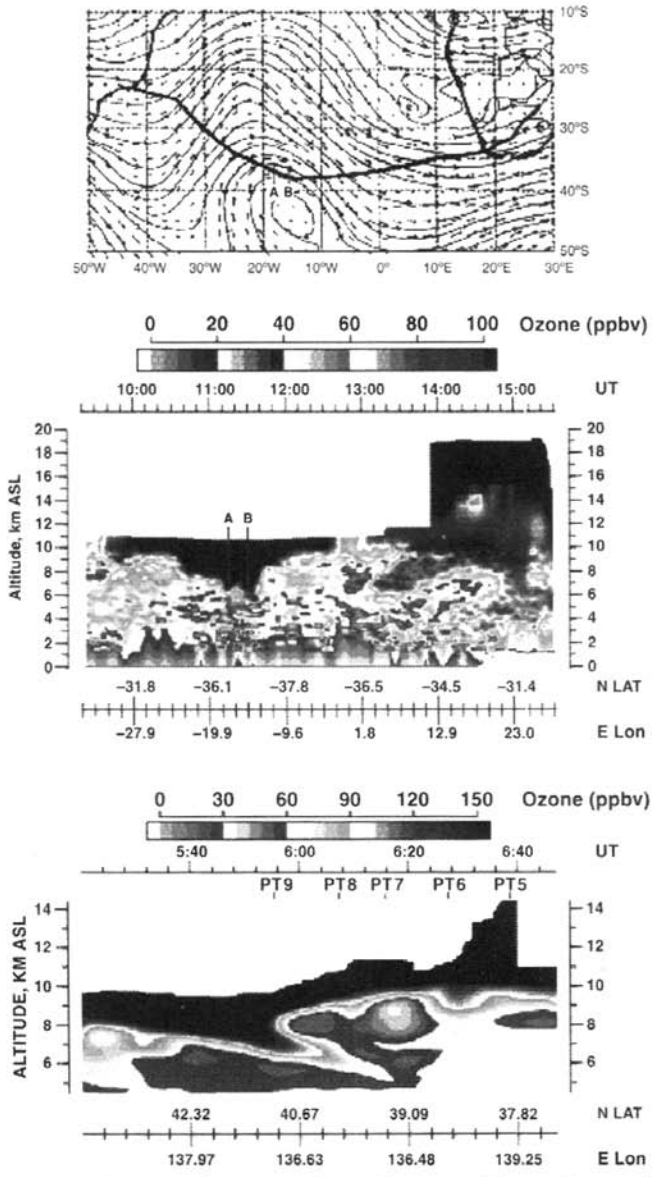


Figure 10 (see color insert) Three-panel figure showing evidence of ozone input from the stratosphere into the troposphere in both hemispheres. The top panel shows the flight path (heavy line) of a DC-8 airplane on October 3, 1992, from South America to Africa that intersected a trough protruding from higher latitudes. Points A and B on that flight path show high concentrations of ozone being transported to altitudes below 6 km in the middle panel; the data depicted in this panel were obtained from a differential absorption lidar system that measured ozone below the 11-km flight level of the DC-8. The lowest panel shows a similar feature for a flight on March 11, 1994, in the Northern Hemisphere. As the airplane flies from north to south in this panel, note the higher tropopause height south of the fold. See ftp site for color image.

systems. Large amounts of stratospheric air extend into the troposphere and much of that air becomes trapped in, and eventually mixed with, the troposphere. An example of stratospheric air coming into the troposphere during a tropopause fold is shown in Figure 10. This figure illustrates the intrusion of stratospheric tracers into the troposphere using a differential absorption laser radar (lidar) instrument that measures ozone below and above it as it flies in an airplane at a cruising altitude of 11 km. The top panel shows the flight path of the airplane (heavy line) and the geopotential height distribution at 200 hPa. This flight path between South America and Africa was part of a field mission in October 1992. Points A and B refer to the location of the two "tongues" of stratospheric air that have descended into the troposphere as the flight path intersected a trough from southern middle latitudes. The middle panel of Figure 10 shows the descent of ozone from the stratosphere (brown areas, >100 ppbv) in conjunction with the tropopause fold (see ftp site for color image). At these points, stratospheric air, as marked by the high concentrations of ozone, has descended to altitudes as low as 6 km. As the flight continues to the east, and as measurements from the upward-looking lidar system became available, the tropopause is located at ~15 km. The higher concentrations of ozone in the middle and upper troposphere (denoted by the orange colors) were formed in situ from widespread biomass burning taking place at this time of the year. The bottom panel is from a flight in March 1994 and perhaps better illustrates the distribution of ozone during a folding event. Note how much higher the tropopause is south of the fold (later in the flight), than at higher latitudes in the beginning of the flight (~8 km at the beginning of the flight), consistent with the schematic shown in Figure 9.

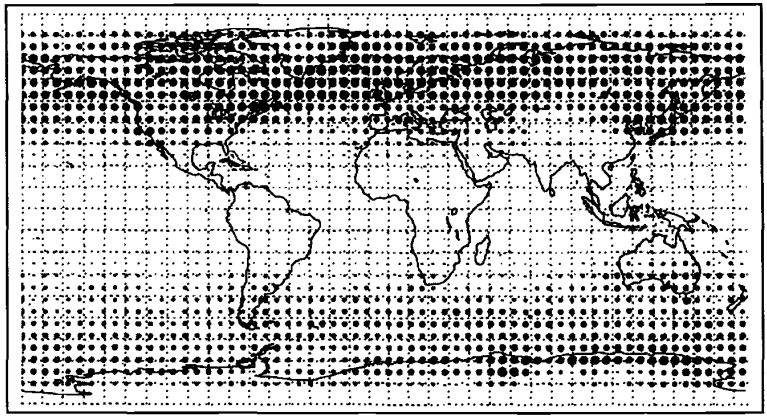


Figure 11 Annual mean distribution of global tropopause folding activity obtained from meteorological analysis over a 10-year period, 1984–1993; the size of the dots denotes the activity corresponding to bringing air from the stratosphere into the troposphere. (Reprinted with permission from Beekman et al., 1997.)

Figure 11 shows the climatological location of stratospheric intrusions weighted by the intensity of the tropopause event to derive a depiction of how much stratospheric air enters the troposphere. The data have been obtained from European Center for Medium-Range Weather Forecasting (ECMWF) data using an identification scheme relating potential vorticity to the exchange of air between the stratosphere and troposphere. This analysis, published in 1997, agrees with previous studies suggesting that considerably more exchange takes place in the Northern Hemisphere relative to the Southern Hemisphere and that the flux in the NH is 6×10^{10} molecules O_3/cm^2 s. This value is in agreement with a number of previous studies since the 1970s that have estimated a cross-tropopause flux using both general circulation models and observations calculating amounts of between 4 and 8×10^{10} molecules O_3/cm^2 s for the NH. With respect to the global tropospheric ozone budget, this "natural" flux of ozone transported would account for only a relatively small fraction of the ozone now commonly measured near Earth's surface, implying that much of the ozone present in the lower atmosphere would not be there without anthropogenic input. The chapter on tropospheric ozone will discuss the *tropospheric ozone* budget in more detail.

REFERENCES

Albritton, D. L., R. T. Watson, and P. J. Aucamp, *Scientific Assessment of Ozone Depletion: 1994*, World Meteorological Organization, Geneva, 1995.

Albritton, D. L., P. J. Aucamp, G. Megie, and R. T. Watson, *Scientific Assessment of Ozone Depletion: 1998*, World Meteorological Organization, Geneva, 1999.

Beekman, M., et al., Regional and global tropopause fold occurrence and related ozone flux across the tropopause, *J. Atmos. Chem.*, *28*, 29–44, 1997.

Brasseur, G., and S. Solomon, *Aeronomy of the Middle Atmosphere* (2nd ed.), Reidel, Dordrecht, 1986.

Farman, J. C., B. G. Gardiner, and J. D. Shanklin, Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction, *Nature*, **315**, 207–210, 1985.

Holton, J. R., A. R. Douglass, P. H. Haynes, M. E. McIntyre, R. B. Rood, and L. Pfister, Stratosphere-troposphere exchange, *Rev. Geophys.*, *33*, 403–439, 1995.

Houghton, J. T., G. J. Jenkins, and J. J. Ephraums (Eds.), Intergovernmental panel on climate change, in *Climate Change, The IPCC Scientific Assessment*, Cambridge University Press, Cambridge, 1990.

Houghton, J. T., L. G. Meira Filho, B. A. Callander, N. Harris, A Kattenberg, and K. Maskell (Eds.), Intergovernmental panel on climate change, in *Climate Change (1995): The Science of Climate Change*, Cambridge University Press, Cambridge, 1996.

Levy II, H., Normal atmosphere: Large radical and formaldehyde concentrations predicted, *Science*, **173**, 141–143, 1971.

Molina, M. J., and F. S. Rowland, Stratospheric sink for chlorofluoromethanes: chlorine catalyzed destruction of ozone, *Nature*, **249**, 810–814, 1974.

Watson, R. T., L. G. Meiro Filho, E. Sanhueza, and A. Jenetos, Greenhouse gases: Sources and sinks, in J. T. Houghton, B. A. Callander, and S. K. Varney (Eds.), *Climate Change 1992: The Supplementary Report to the IPCC Scientific Assessment*, Cambridge University Press, Cambridge, 1992, pp.1–40.

CHAPTER 2

OXIDIZING POWER OF ATMOSPHERE

DANIEL J. JACOB

1 INTRODUCTION

The atmosphere is an oxidizing medium. Many environmentally important trace gases are removed from the atmosphere by oxidation, including methane and other organic compounds, carbon monoxide, nitrogen oxides, and sulfur gases (Table 1). Understanding the processes and rates by which species are oxidized in the atmosphere, i.e., the oxidizing power of the atmosphere, is crucial to our knowledge of atmospheric composition. Changes in the oxidizing power of the atmosphere would have a wide range of implications for air pollution, aerosol formation, greenhouse radiative forcing, and stratospheric ozone depletion (Thompson, 1992).

The most abundant oxidants in Earth's atmosphere are O_2 and O_3. They have large bond energies and are hence relatively unreactive. With a few exceptions, oxidation of nonradical atmospheric species by O_2 or O_3 is negligibly slow. Photochemical modeling of stratospheric chemistry in the 1950s first implicated the strong radical oxidants O and OH, generated from photolysis of O_3 and H_2O, in the oxidation of CO and CH_4 (Bates and Witherspoon, 1952). The importance of photochemically generated radicals in the chain oxidation of hydrocarbons leading to urban O_3 smog was also recognized in the 1950s (Leighton, 1961). Smog models of that time hypothesized that O atoms produced in urban air from the photolysis of NO_2 and O_3 would provide the main pathway for hydrocarbon oxidation (Altshuller and Bufalini, 1965, 1971). This mechanism was thought unimportant outside of urban areas because of low O_3 and NO_2 concentrations, and transport to the stratosphere was viewed as necessary for oxidation of CO, CH_4, and other gases present in the global troposphere (Cadle and Allen, 1970). Long atmospheric lifetimes for these gases were implied because of the 10-year residence time of air in the troposphere.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

TABLE 1 Atmospheric Lifetimes of Selected Species

Species	Lifetime[a]	Reference
CH_3CCl_3	4.8 yr (*5.7 yr*)	WMO (1999)
CH_4	8.4 yr (*8.9 yr*)	WMO (1999)
CHF_2Cl	11.8 yr (*12.3 yr*)	WMO (1999)
CH_3Br	0.7 yr (*1.7 yr*)	WMO (1999)
Isoprene[b]	~ 1 h (~ *1 h*)	Jacob et al. (1989)
CO	2 mo (*2 mo*)	Logan et al. (1981)
NO_x ($NO + NO_2$)	~ 1 d (~ *1 d*)[c]	Dentener and Crutzen (1993)
SO_2	~ 1 d (*2 wks*)[d]	Chin et al. (1996)
$(CH_3)_2S$	~ 1 d (~ *1 d*)	Chin et al. (1996)

[a]The atmospheric lifetime of a species is defined as the average time that a molecule of the species remains in the atmosphere before it is removed by one of its sinks. It can be calculated as the atmospheric mass of the species divided by the species loss rate. The first number given for each entry in the column is the mean atmospheric lifetime, and the second number in parentheses is the mean atmospheric lifetime against oxidation by OH.

[b]$CH = C(CH_3)-CH = CH_2$, a major hydrocarbon emitted by vegetation.

[c]Loss of NO_x in summer and in the tropics is mostly by reaction of NO_2 with OH; loss in winter at extratropical latitudes is mostly by a nonphotochemical pathway involving formation of N_2O_5 and hydrolysis to HNO_3. The sum of these two processes results in a lifetime of NO_x of the order of a day.

[d]The principal SO_2 sinks are deposition and in-cloud oxidation by $H_2O_2(aq)$.

This view of a chemically inert troposphere was first challenged by Weinstock (1969) who found from ^{14}CO measurements that the atmospheric lifetime of CO is only ~0.1 years, requiring a dominant sink in the troposphere. Levy (1971) then presented photochemical model calculations for the unpolluted troposphere showing that high concentrations of OH could be generated from photolysis of O_3 in the presence of water vapor and account for the missing sink of CO in the Weinstock (1969) analysis. Further work in the early 1970s confirmed the importance of tropospheric oxidation by OH as the main sink of CO and CH_4 (McConnell et al., 1971; Weinstock and Niki, 1972; Levy et al., 1973) and further showed that OH, not O, is the main oxidant of hydrocarbons in urban air (Heicklen, 1971; Kerr et al., 1972; Demerjian et al., 1974). Considerable evidence over the past three decades supports the view that tropospheric OH is the main oxidant for nonradical species in the atmosphere.

Indirect estimates of global mean OH concentrations have been made since the 1970s using a number of proxies, the most useful of which has been CH_3CCl_3, a long-lived gas emitted by industry and removed from the atmosphere by oxidation by OH (Lovelock, 1977; Singh, 1977). The most recent analyses of CH_3CCl_3 data, based on observations at a worldwide network of sites (Prinn et al., 1995), imply a global mean OH concentration in the troposphere of $(1.1 \pm 0.1) \times 10^6$ molecules/cm^3 (Krol et al., 1998; Spivakovsky et al., 2000). Techniques for direct measurement of tropospheric OH were first developed in the 1970s but suffered from

interferences or poor sensitivity. Only in the 1990s have reliable techniques been developed and successfully intercompared (special issue of *Journal of the Atmospheric Sciences*, October 1995; Crosley, 1997). Direct measurements provide the means to test our understanding of the local processes controlling OH concentrations (e.g., McKeen et al., 1997; Jaeglé et al., 1997, 2000; Frost et al., 1999). By simulating these processes in global models, one can assess the sensitivity of the oxidizing power of the atmosphere to different anthropogenic perturbations (Wang and Jacob, 1998).

This chapter reviews current understanding of the factors controlling abundances and long-term trends of OH. It also briefly reviews (Section 3) other atmospheric oxidants that are important in certain environments or for certain nonradical molecules. It does not cover the oxidation of short-lived radical species, which often involves reaction with O_2 or O_3 (Atkinson, 1990). It does not cover either oxidation in the stratosphere, whose importance as a sink for species emitted at the surface is limited by the long time for transfer of air from the troposphere to the stratosphere.

2 HYDROXYL RADICAL OH

Processes Controlling OH Concentrations

A detailed and still fairly current discussion of OH chemistry in the troposphere is given by Logan et al. (1981). The primary source of OH is the photolysis of O_3 to produce an excited state of atomic oxygen, $O(^1D)$, which then reacts with water vapor:

$$O_3 + h\nu \rightarrow O_2 + O(^1D) \qquad (\lambda < 340 \text{ nm}) \qquad \text{(R1)}$$

$$O(^1D) + M \rightarrow O(^3P) + M \qquad \text{(R2)}$$

$$O(^3P) + O_2 + M \rightarrow O_3 + M \qquad \text{(R3)}$$

$$H_2O + O(^1D) \rightarrow 2OH \qquad \text{(R4)}$$

Here M is an inert molecule (N_2 or O_2). Only $\sim 1\%$ of the $O(^1D)$ atoms produced by (R1) react with H_2O; most are deactivated to the ground-state $O(^3P)$ and recombine with O_2 to return O_3. Photolysis of O_3 to $O(^1D)$ in the troposphere is determined by a narrow band of radiation in the 290- to 330-nm range, reflecting the combined wavelength dependences of the actinic flux, O_3 absorption cross section, and $O(^1D)$ quantum yield (Fig. 1). Radiation in this wavelength range is strongly absorbed by overhead O_3, and hence the production of $O(^1D)$ is strongly dependent on the thickness of the stratospheric O_3 layer (Madronich and Granier, 1992).

The OH radical is consumed on a time scale of ~ 1 s by oxidation of a large number of reduced atmospheric species. Its main sinks in the troposphere are CO and CH_4. Nonmethane hydrocarbons (NMHCs) are also important sinks in the lower troposphere over continents. Oxidation of CO or hydrocarbons by OH propagates a

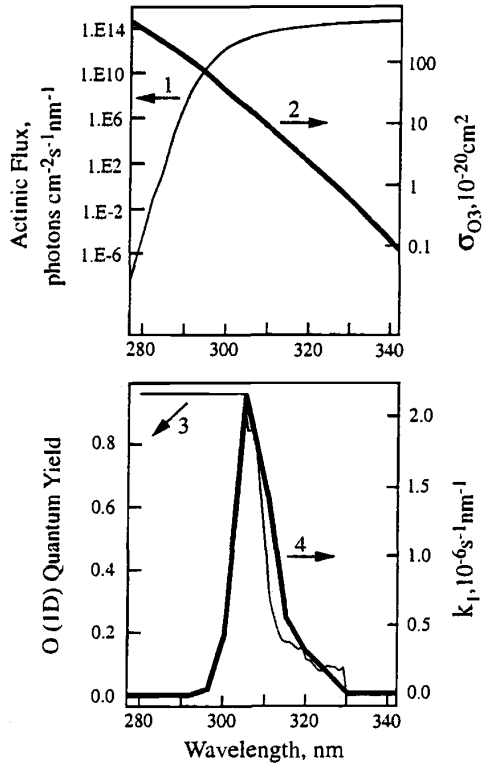


Figure 1 Computation of the rate constant k_1 of reaction (R1) as the integral over all wavelengths of the actinic flux of solar radiation (1) times the absorption cross-section σ_{03} of ozone (2) and times the O(ID) quantum yield (3). From Jacob (1999).

radical reaction chain initiated by the generation of OH radicals from (R4). The simplest case is oxidation of CO:

$$CO + OH \rightarrow CO_2 + H \quad \text{(R5)}$$

$$H + O_2 + M \rightarrow HO_2 + M \quad \text{(R6)}$$

The HO_2 radicals may self-react to produce H_2O_2 (hydrogen peroxide):

$$HO_2 + HO_2 \rightarrow H_2O_2 + O_2 \quad \text{(R7)}$$

or they may regenerate OH by reaction with NO or O_3:

$$HO_2 + NO \rightarrow OH + NO_2 \quad \text{(R8)}$$

$$HO_2 + O_3 \rightarrow OH + 2O_2 \quad \text{(R9)}$$

Hydrogen peroxide produced by (R7) is removed from the atmosphere by deposition. It may also photolyze, regenerating OH,

$$H_2O_2 + h\nu \rightarrow 2OH \qquad \text{(R10)}$$

or react itself with OH:

$$H_2O_2 + OH \rightarrow H_2O + HO_2 \qquad \text{(R11)}$$

The same type of chain mechanism applies to the oxidation of hydrocarbons, but the complexity increases rapidly as the size of the hydrocarbon molecule increases. The mechanism for CH_4 is described here. It begins by

$$CH_4 + OH \rightarrow CH_3 + H_2O \qquad \text{(R12)}$$
$$CH_3 + O_2 + M \rightarrow CH_3O_2 + M \qquad \text{(R13)}$$

The CH_3O_2 molecule (methylperoxy radical) is analogous to HO_2. Its dominant sinks in the atmosphere are reactions with HO_2 and NO:

$$CH_3O_2 + HO_2 \rightarrow CH_3OOH + O_2 \qquad \text{(R14)}$$
$$CH_3O_2 + NO \rightarrow CH_3O + NO_2 \qquad \text{(R15)}$$

Similarly to H_2O_2, methylhydroperoxide (CH_3OOH) may either react with OH or photolyze:

$$CH_3OOH + OH \rightarrow CH_2O + OH + H_2O \qquad \text{(R16)}$$
$$CH_3OOH + OH \rightarrow CH_3O_2 + H_2O \qquad \text{(R17)}$$
$$CH_3OOH + h\nu \rightarrow CH_3O + OH \qquad \text{(R18)}$$

The methoxy radical CH_3O produced by (R15) and (R18) reacts rapidly with O_2:

$$CH_3O + O_2 \rightarrow CH_2O + HO_2 \qquad \text{(R19)}$$

Formaldehyde produced by (R16) and (R19) may either react with OH or photolyze (two photolysis branches):

$$CH_2O + OH \longrightarrow CHO + H_2O \qquad \text{(R20)}$$
$$CH_2O + h\nu \xrightarrow{O_2} CHO + HO_2 \qquad \text{(R21)}$$
$$CH_2O + h\nu \longrightarrow CO + H_2 \qquad \text{(R22)}$$

Reactions (R20) and (R21) produce the CHO radical, which reacts rapidly with O_2 to yield CO:

$$CHO + O_2 \rightarrow CO + HO_2 \qquad \text{(R23)}$$

In this overall sequence the C(−IV) atom in CH_4 is gradually oxidized to C(−II) in CH_3OOH, C(0) in CH_2O, C(+II) in CO, and C(+IV) in CO_2 (highest oxidation state for carbon).

The regeneration of OH radicals by (R8) plays a critical role in maintaining OH concentrations in the troposphere. The main sink for NO_2 produced by (R8) and (R15) is photolysis, regenerating NO and producing O_3:

$$NO_2 + h\nu \rightarrow NO + O(^3P) \qquad \text{(R24)}$$

$$O(^3P) + O_2 + M \rightarrow O_3 + M \qquad \text{(R3)}$$

This O_3 may then photolyze to yield additional OH by (R1) + (R4). Although reaction (R9) also recycles OH, it consumes in the process an O_3 molecule that could have otherwise photolyzed to produce OH. Therefore, it is not effective for maintaining OH concentrations.

Figure 2 illustrates how tropospheric OH is controlled by chemical cycling of the hydrogen oxide family ($HO_x \equiv$ OH+ peroxy radicals) and the nitrogen oxide family ($NO_x \equiv NO + NO_2$), for the simple case of CO oxidation. The schematic for hydrocarbon oxidation is similar, except that photolysis of carbonyl compounds as in reaction (R21) provides an additional (generally minor) source of HO_x. The dominant sink for the HO_x family is usually the formation of peroxides. As discussed

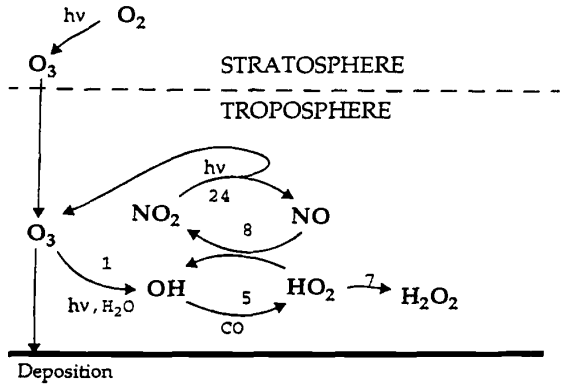


Figure 2 Simplified schematic of O_3–HO_x–NO_x–CO chemistry in the troposphere.

previously, these peroxides may photolyze to recycle HO_x; alternatively, they may deposit or react with OH, providing a terminal sink for HO_x. Sources of NO_x in the troposphere include combustion, microbial activity in soils, and lightning. Sources of CO and hydrocarbons include combustion, industrial processes, soils, and vegetation.

An analytical expression for the dependence of OH concentrations on chemical variables can be obtained from the simplified O_3–HO_x–NO_x–CO system by assuming chemical steady state for the short-lived species $O(^1D)$, H, OH, and also for the chemical family HO_x. The lifetime of HO_x against formation of peroxides is of the order of minutes, so that the steady-state assumption is appropriate. The production rate P_{HO_x} of HO_x from reaction (R4) is given by

$$P_{HO_x} = 2k_4[O(^1D)][H_2O] \equiv 2\frac{k_1 k_4}{k_2[M]}[O_3][H_2O] \tag{1}$$

where k_i is the rate constant for reaction i. In writing Eq. (1) we have used the approximation (R2)$\gg$(R4) to simplify the denominator. Steady state for OH is defined by

$$P_{HO_x} + k_8[HO_2][NO] = k_5[CO][OH] \tag{2}$$

Loss of HO_x in this system is by (R7). Steady state for HO_x is therefore defined by

$$P_{HO_x} = 2k_7[HO_2]^2 \tag{3}$$

from which we derive the following expression for the OH concentration:

$$[OH] = \frac{P_{HO_x} + k_8\sqrt{\frac{P_{HO_x}}{2k_7}}[NO]}{k_5[CO]} \tag{4}$$

We see from Eq. (4) together with Eq. (1) that OH concentrations depend negatively on CO and positively on water vapor, O_3, and NO. The dependence on hydrocarbons is more complicated (as hydrocarbons provide both sinks of OH and sources of HO_x) but is generally negative, similar to CO.

One important caveat to this simplified representation of OH chemistry must be made for high-NO_x environments. When NO_x concentrations exceed a few parts per billion by volume (ppbv), as in urban air, oxidation of NO_2 by OH can become the dominant sink for HO_x:

$$NO_2 + OH + M \rightarrow HNO_3 + M \tag{R25}$$

Under these conditions, OH concentrations decrease with increasing NO_x (as may be derived by repeating the steady-state calculation above) and increase with increasing

hydrocarbons. This situation is commonly denoted the NO_x-saturated (or hydrocarbon-limited) regime, as opposed to the NO_x-limited regime normally encountered in the troposphere.

A second caveat applies to the upper troposphere where water vapor concentrations are low (~100 ppmv). Under these conditions, reaction (R4) may be less important as a primary source of HO_x than photolysis of acetone originating from the biosphere (Singh et al., 1995) or convective injection of peroxides and aldehydes produced in the lower troposphere (Jaeglé et al., 1997; Prather and Jacob, 1997; Müller and Brasseur, 1999). Reaction of OH with HO_2 provides in general the dominant HO_x sink in the upper troposphere, which yields a square root rather than linear dependence of OH concentrations on NO.

Figure 3 shows zonal mean global distributions of OH concentrations computed with a global three-dimensional model of tropospheric O_3–NO_x–hydrocarbon chemistry (Wang et al., 1998b). The highest concentrations (averaging over 2×10^6 molecules/cm^3) are in the tropical middle troposphere, reflecting a combination of high ultraviolet (UV) and high humidity. The large seasonal variation at midlatitudes follows UV radiation. Concentrations tend to be higher in the Northern than in the Southern Hemisphere, reflecting higher NO_x concentrations.

Global Mean OH Concentration

The short lifetime of OH implies that its concentration is highly variable. Deriving the atmospheric lifetimes of gases removed by oxidation by OH requires an estimate of OH concentrations averaged appropriately over time and space. Mass-balance arguments for proxy species with known sources can assist for this purpose. The most successful application, first proposed by Singh (1977) and Lovelock (1977), has been the use of the industrial solvent CH_3CCl_3 to estimate the global mean OH concentration. The source of CH_3CCl_3 is exclusely anthropogenic, and its historical trend is well known from industrial data. Production of CH_3CCl_3 has been banned since 1996 as part of the Montreal Protocol. The dominant sink of CH_3CCl_3 is oxidation by OH in the troposphere (photolysis in the stratosphere and uptake by the oceans are small additional sinks). Tropospheric mixing ratios of CH_3CCl_3 are relatively uniform, so that a mass-balance analysis for CH_3CCl_3 yields a global mean OH concentration weighted by atmospheric mass and by the temperature dependence of the $CH_3CCl_3 + OH$ reaction. The global mean OH concentration obtained in this manner can then be used to infer the lifetimes of other long-lived gases removed by reaction with OH, such as CH_4 and hydrogenated halocarbons (HCFCs) (Prather and Spivakovsky, 1990).

The most recent use of CH_3CCl_3 observations to constrain the global mean OH concentration has been by Krol et al. (1998) and Spivakovsky et al. (2000). These authors derive a CH_3CCl_3 lifetime of 5.5 years in the troposphere against oxidation by OH, corresponding to a global mean OH concentration of $(1.1 \pm 0.2) \times 10^6$ molecules/cm^3. Spivakovsky et al. (2000) point out that the magnitude of the CH_3CCl_3 interhemispheric gradient implies that the difference between the mean OH concentrations in the Northern and Southern Hemispheres is no more than 50%.

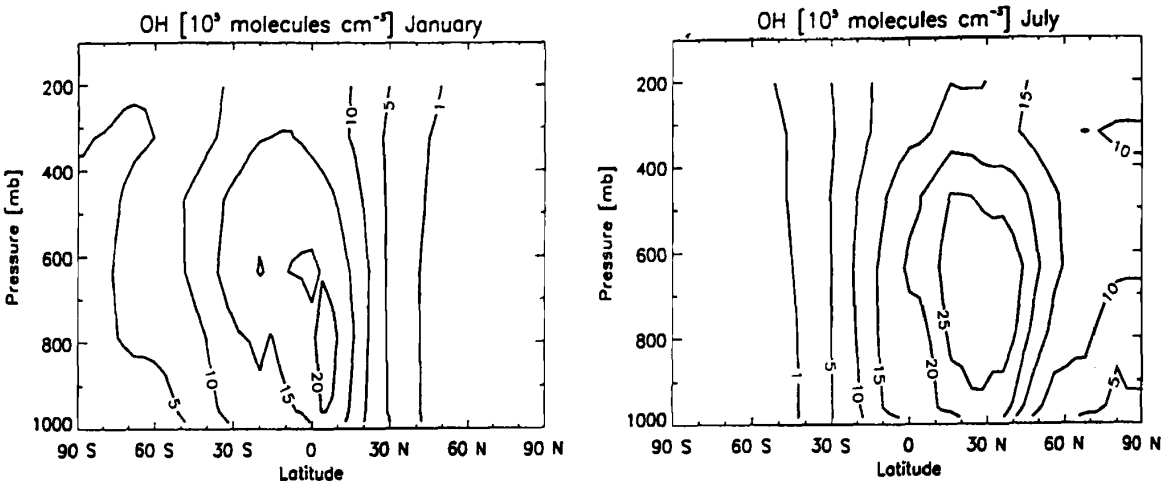


Figure 3 Longitudinally averaged monthly mean OH concentrations.

Mass-balance arguments for other chemical tracers oxidized by OH including ^{14}CO, CHF_2Cl, CH_2Cl_2, and hydrocarbons have been used to confirm the above estimate of the global mean OH concentration and to provide additional constraints on the geographical and seasonal distribution of OH (Volz et al., 1981; Mak et al., 1992; Goldstein et al., 1995; Spivakovsky et al., 2000).

Simulation of the CH_3CCl_3 lifetime has long been a standard test for evaluating the global mean OH concentration computed in tropospheric chemistry models, starting from the work of Crutzen and Fishman (1977). In these models, the OH concentrations are computed from a global simulation of O_3–NO_x–CO–hydrocarbon chemistry that treats emissions, transport, chemistry, and deposition in a self-consistent way (e.g., Wang et al., 1998a). The current generation of models reproduces the atmospheric lifetime of CH_3CCl_3 to within typically 25%.

Measurements of OH Concentrations and Comparisons to Models

The past few years have seen the development of a number of methods for direct measurement of tropospheric OH (special issue of *Journal of Atmospheric Science*, October 1995). Two of these methods, a long-path absorption (LPA) instrument (Mount, 1992) and a chemical ionization mass spectrometry (CIMS) instrument (Eisele and Tanner, 1991) were intercompared formally at a mountain site in Colorado during the Tropospheric OH Photochemistry Experiment (TOHPE). Under well-mixed atmospheric conditions where the local OH measurement from CIMS could be compared to the long-path average from LPA, the intercomparison demonstrated a good correlation between the two instruments down to concentrations of less than 1×10^6 molecules/cm^3, with no significant bias (Crosley, 1997).

A number of ancillary chemical measurements were made during TOHPE that McKeen et al. (1997) used to compare the observed OH concentrations to values computed from a standard photochemical model. The model overestimated OH concentrations by a factor of 1.3 on average. It captured 48% of the variance in the CIMS instrument, although much of that variance was driven by the diurnal cycle. It was not correlated with the LPA instrument, which may reflect the nonlocal nature of the latter measurement.

The model overestimate of OH in TOHPE is consistent with other model measurement comparisons conducted at continental sites (Poppe et al., 1995; Thompson, 1995; George et al., 1999). As discussed by McKeen et al. (1997), possible causes include inadequate model representation of hydrocarbon chemistry or of uptake of HO_x by aerosols. Eisele et al. (1996) conducted a model-measurement comparison using the CIMS instrument at Mauna Loa Observatory, Hawaii (3.4 km altitude); they found good agreement when subsiding motions brought free tropospheric air to the site but a factor of 2 model overestimate under upslope flow, supporting the view that biogenic hydrocarbons may provide important sinks for OH. Frost et al. (1999) found a median model overestimate of 32% in simulation of aircraft observations for clean marine air.

An important aspect of these model-measurement comparisons has been to examine the ability of models to reproduce the dependence of OH concentrations on chemical and meteorological variables. Poppe et al. (1995) found that their model

could capture successfully the observed correlations of OH concentrations with UV intensity, temperature, humidity, and CO concentration. Measurements in TOHPE showed OH concentrations increasing with increasing NO_x up to about 2 ppbv NO_x and then decreasing, consistent with model calculations of NO_x versus hydrocarbon-limited chemistry (Eisele et al., 1997; McKeen et al., 1997).

Aircraft measurements of OH and HO_2 concentrations in the upper troposphere have been reported by Brune et al. (1998, 1999) and Wennberg et al. (1998). The measured OH/HO_2 ratios and their variances agree with model values to within the uncertainties of the relevant rate constants, implying a good understanding of the cycling of HO_x (Jaeglé et al., 2000). The observed HO_x concentrations are often several times lower than would be predicted solely from the $O(^1D) + H_2O$ source (R4) and support the presence of other primary HO_x sources in the upper troposphere including acetone, peroxides, and aldehydes.

Long-Term Trends in Atmospheric OH

Assessing human influence on the oxidizing power of the atmosphere is intricate. On the one hand, anthropogenic emissions of CO and hydrocarbon emissions act to deplete OH; on the other hand, anthropogenic emissions of NO_x and the thinning of the stratospheric O_3 layer act to boost OH. Human-induced changes in Earth's climate (temperature, cloudiness, circulation) add to the complication. Large regional differences may be expected in the response of OH to human activity, depending on the relative importance and coupling of the above factors.

A number of global tropospheric chemistry model studies, reviewed by Thompson (1992), have examined the changes in OH concentrations since preindustrial times as driven by trends in emissions of CO, hydrocarbons, and NO_x. These studies report 10 to 30% decrease in the global mean OH concentration from preindustrial times to today, a relatively small effect considering that emissions of CO, CH_4, and NO_x increased severalfold over that period (Table 2). The global three-dimensional model study of Wang and Jacob (1998) indicate a 9% decrease in the global mean

TABLE 2 Comparison of Present and Preindustrial Atmospheres[a]

	Emission					
	CH_4 (Tg CH_4/yr)	Nonmethane Hydrocarbons (Tg C/yr)	CO (Tg CO/yr)	NO_x (Tg N/yr)	O_3 Source[b] (Tg O_3/yr)	[OH][c] (molecules/cm^3)
Preindustrial	160	610[d]	50	9	2300	1.15×10^6
Present	460	710	1040	42	4500	1.04×10^6

[a]Global data from the three-dimensional model study of Wang and Jacob (1998).
[b]Tropospheric O_3 source including transport from the stratosphere (400 Tg O_3/yr in both preindustrial and present cases) and chemical production within the troposphere.
[c]Global mean tropospheric concentration weighted by atmospheric mass.
[d]Biogenic isoprene and acetone.

OH concentration since preindustrial times and suggests that the OH trend should follow roughly the trend of the $S_{NO}/S_C^{3/2}$ ratio, where S_{NO} is the global source of NO and S_C is the global source of CO and hydrocarbons; the parallel changes in S_{NO} and S_C over the past century would thus have had nearly cancelling effects on OH concentrations. This study points out that estimates of past and future trends in OH are highly sensitive to assumed trends in tropical biomass burning because NO_x emitted in the tropics is particularly efficient for generating O_3 and OH.

Observational constraints on long-term OH trends are largely limited to the CH_3CCl_3 record since 1978. An analysis of this record by Krol et al. (1998) indicates a 0.5%/yr increase in global mean OH concentrations over the period 1978 to 1993. This result is consistent with radiative transfer model calculations by Madronich and Granier (1992), which indicate a 0.4%/yr increase in OH concentrations over the 1979 to 1989 decade as a result of stratospheric O_3 depletion.

Estimates of OH trends since preindustrial and glacial times have been made using polar ice core records of CH_2O and H_2O_2. Interpretation of these records is complicated by postdepositional exchange with the atmosphere and reactions within the ice (Neftel et al., 1995). Also, since CH_2O and H_2O_2 have atmospheric lifetimes of about a day, they can only diagnose trends in polar OH, which may be different from global tropospheric trends. Analysis of the CH_2O/CH_4 ratio in a Greenland ice core (Staffelbach et al., 1991) suggests that OH concentrations were 30% higher in the preindustrial atmosphere than today, and 2 to 4 times lower in the last glacial maximum (LGM) than today. Such depletion of OH in the LGM is not consistent with results from tropospheric chemistry models, which indicate higher OH concentrations in glacial than interglacial periods due to lower emissions of CH_4 (Thompson, 1992). Staffelbach et al. (1991) suggested that a thicker stratospheric O_3 layer could be responsible for low OH levels during glacial periods.

Data for H_2O_2 in Greenland ice going back to A.D. 1300 show constant concentrations until about 1970, and a doubling of concentrations since then (Sigg and Neftel, 1991; Anklin and Bales, 1997). Although the rise in H_2O_2 would imply a rise in HO_x, the CH_3CCl_3 record shows no large trends in global mean OH concentrations during that same period.

3 OTHER ATMOSPHERIC OXIDANTS

Other atmospheric oxidants besides OH may also be important in some environments and for some species. They are reviewed briefly below.

Nitrate Radical

The nitrate radical (NO_3) is a strong radical oxidant formed in the oxidation of NO_2 by O_3:

$$NO_2 + O_3 \rightarrow NO_3 + O_2 \qquad \text{(R26)}$$

A detailed review of its atmospheric chemistry is given by Wayne (1991). During the daytime, NO_3 photolyzes on a time scale of 1 min to return NO_2:

$$NO_3 + h\nu \rightarrow NO_2 + O \qquad \text{(R27)}$$

At night the lifetime of NO_3 is much longer. In high-NO_x regions such as the eastern United States, NO_3 accumulates to concentrations of 10 to 100 parts per trillion by volume (pptv) during the nighttime hours (Wayne, 1991). At these concentrations, NO_3 can provide an important sink for some unsaturated hydrocarbons including isoprene and terpenes (O_3 is also an important oxidant for these compounds). Measurements in relatively polluted marine air over the North Sea indicate a mean nighttime NO_3 concentration of about 10 pptv; at this concentration, NO_3 represents a major sink for biogenic dimethylsulfide (Carslaw et al., 1997). Nighttime accumulation of NO_3 is in general limited by equilibrium with N_2O_5, followed by hydrolysis of N_2O_5 in aerosols:

$$NO_3 + NO_2 + M \longrightarrow N_2O_5 + M \qquad \text{(R28)}$$

$$N_2O_5 \xrightarrow{\text{heat}} NO_3 + NO_2 \qquad \text{(R29)}$$

$$N_2O_5 + H_2O \xrightarrow{\text{aerosol}} 2HNO_3 \qquad \text{(R30)}$$

At low temperatures ($T < 280$ K) the NO_3/N_2O_5 equilibrium is shifted far to the right; thus NO_3 is important only in the warm lower troposphere.

Halogen Radical Oxidants

There has been longstanding interest in the possible role of halogen radicals as tropospheric oxidants (Singh and Kasting, 1988; Chatfield and Crutzen, 1990). The best evidence so far comes from measurements of alkanes and acetylene in Arctic surface air (Jobson et al., 1994), which indicate a sink in April (polar sunrise) consistent with oxidation by Cl atoms present at a concentration of $\sim 1 \times 10^4$ atoms/cm^3. The data also suggest the presence of Br atoms to oxidize acetylene. The source of the halogen oxidants is not well established but likely involves chemical production from sea salt accumulated on the ice over the polar night (Impey et al., 1999).

Generation of halogen oxidants from sea salt would be of little interest for global tropospheric chemistry if it were confined to Arctic sunrise. However, measurements of hydrocarbons and nonradical Cl species in the marine boundary layer (MBL) at midlatitudes and in the tropics suggest that Cl atoms may be present at least occasionally at concentrations in the range 10^4 to 10^5 atoms/cm^3 (Keene et al., 1990, 1996; Pszenny et al., 1993; Singh et al., 1996; Spicer et al., 1998). At such concentrations, oxidation by Cl atoms would provide a major sink for dimethylsulfide and alkanes in the MBL. Even less is known about Br radical chemistry in the MBL, although Toumi (1994) has suggested that BrO could provide an important oxidant

for dimethylsulfide. Field measurements of the halogen radicals and their reservoirs HOCl and HOBr are needed.

Cloud and Aerosol Oxidants

Water-soluble atmospheric species incorporated in cloud droplets and aqueous aerosols may dissociate into ions, and the resulting aqueous-phase redox chemistry provides yet another pathway for oxidation of species in the atmosphere. The importance of this pathway has been established for SO_2, which dissociates in water to HSO_3^- and SO_3^{2-} ($pK_{a1} = 1.9$, $pK_{a2} = 7.2$). Rapid oxidation of SO_2 by H_2O_2 in cloud was first suggested by Penkett et al. (1979):

$$SO_2(g) \Leftrightarrow SO_2 \cdot H_2O \quad \text{(R31)}$$

$$SO_2 \cdot H_2O \Leftrightarrow HSO_3^- + H^+ \quad \text{(R32)}$$

$$H_2O_2(g) \Leftrightarrow H_2O_2(aq) \quad \text{(R33)}$$

$$HSO_3^- + H_2O_2(aq) + H^+ \rightarrow SO_4^{2-} + 2H^+ + H_2O \quad \text{(R34)}$$

Aircraft measurements by Daum et al. (1984) demonstrated that the reaction is sufficiently fast to titrate either SO_2 or H_2O_2 in cloud (whichever is limiting). It is now well accepted that this mechanism dominates over gas-phase oxidation by OH as a sink for SO_2 in the atmosphere (Chin et al., 1996).

Additional nonradical oxidants may also be important for oxidation of SO_2 in clouds and aqueous aerosols, but their importance is not as well verified as for H_2O_2. At high pH values (pH > 5), $O_3(aq)$ reacts rapidly with SO_3^{2-}:

$$HSO_3^- \Leftrightarrow SO_3^{2-} + H^+ \quad \text{(R35)}$$

$$SO_3^{2-} + O_3(aq) \rightarrow SO_4^{2-} + O_2 \quad \text{(R36)}$$

This mechanism, taking place in alkaline sea salt aerosols, could represent a major sink for SO_2 in the marine boundary layer (Chameides and Stelson, 1992). Additional SO_2 oxidants in sea salt aerosol may include HOCl and HOBr produced by halogen radical chemistry (Vogt et al., 1996). In polluted clouds, aqueous-phase autoxidation catalyzed by Fe(III) could provide the dominant SO_2 sink (Jacob and Hoffmann, 1983).

ACKNOWLEDGMENT

I thank Clarisa M. Spivakovsky for valuable comments.

REFERENCES

Altshuller, A. P., and J. Bufalini, Photochemical aspects of air pollution: A review, *Photochem. Photobiol.*, *4*, 97–146, 1965.

Altshuller, A. P., and J. Bufalini, Photochemical aspects of air pollution: A review, *Environ. Sci. Technol.*, *5*, 39–62, 1971.

Anklin, M., and R. C. Bales, Recent increases in H_2O_2 concentrations at Summit, Greenland, *J. Geophys. Res.*, *102*, 19099–19104, 1997.

Atkinson, R. A., Gas-phase tropospheric chemistry of organic compounds: A review, *Atmos. Environ.*, *24*, 1–42, 1990.

Bates, D. R., and A. Witherspoon, The photochemistry of some minor constituents of the earth's atmosphere (CO_2, CO, CH_4, N_2O), *Mon. Not. Roy. Astron. Soc.*, *112*, 101, 1952.

Brune, W. H., et al., Airborne in-situ OH and HO_2 observations in the cloud-free troposphere and lower stratosphere during SUCCESS, *Geophys. Res. Lett.*, *25*, 1701–1704, 1998.

Brune, W. H., et al., OH and HO_2 chemistry in the North Atlantic free troposphere, *Geophys. Res. Lett.*, *26*, 3077–3080, 1999.

Cadle, R. D., and E. R. Allen, Atmospheric photochemistry, *Science*, *167*, 243–249, 1970.

Carslaw, N., L. J. Carpenter, J. M. C. Plane, B. J. Allan, R. A. Burgess, K. C. Clemitshaw, H. Coe, and S. A. Penkett, Simultaneous observations of nitrate and peroxy radicals in the marine boundary layer, *J. Geophys. Res.*, *102*, 18917–18933, 1997.

Chameides, W. L., and A. W. Stelson, Aqueous-phase chemical processes in deliquescent sea-salt aerosols: A mechanism that couples the atmospheric cycles of S and sea salt, *J. Geophys. Res.*, *97*, 20565–20580, 1992.

Chatfield, R. C., and P. J. Crutzen, Are there interactions of iodine and sulfur species in marine air photochemistry? *J. Geophys. Res.*, *95*, 22319–22342, 1990.

Chin, M., D. J. Jacob, G. M. Gardner, M. S. Foreman-Fowler, and P. A. Spiro, A global three-dimensional model of tropospheric sulfate, *J. Geophys. Res.*, *101*, 18667–18690, 1996.

Crosley, D. R., The measurement of OH and HO_2 in the troposphere, *J. Atmos. Sci.*, *52*, 3299–3314, 1995.

Crosley, D. R., 1993 Trospospheric OH Experiment: A summary and perspective, *J. Geophys. Res.*, *102*, 6495–6510, 1997.

Crutzen, P. J., and J. Fishman, Average concentrations of OH in the troposphere, and the budgets of CH_4, CO, H_2 and CH_3CCl_3, *Geophys. Res. Lett.*, *4*, 321–324, 1977.

Daum, P. H., S. E. Schwartz, and L. Newman, Acidic and related constituents in liquid-water clouds, *J. Geophys. Res.*, *89*, 1447–1458, 1984.

Demerjian, K. L., J. A. Kerr, and J. G. Calvert, The mechanism of photochemical smog formation, *Adv. Environ. Sci. Technol.*, *4*, 1–262, 1974.

DeMore, W. B., S. P. Sander, D. M. Golden, R. F. Hampson, M. J. Kurylo, C. J. Howard, A. R. Ravishankara, C. E. Kolb, and M. J. Molina, Chemical kinetics and photochemical data for use in stratospheric modeling, *JPL Publication 97–4*, Pasadena, CA, 1997.

Dentener, F. J., and P. J. Crutzen, Reaction of N_2O_5 on tropospheric aerosols: Impact on the global distributions of NO_x, O_3, and OH, *J. Geophys. Res.*, *98*, 7149–7163, 1993.

Eisele, F. L., G. H. Mount, D. Tanner, A. Jefferson, R. Shetter, J. W. Harder, and E. J. Williams, Understanding the production and interconversion of the hydroxyl radical during the Tropospheric OH Photochemistry Experiment, *J. Geophys. Res.*, *102*, 6457–6465, 1997.

Eisele, F. L., and D. J. Tanner, Ion-assisted tropospheric OH measurements, *J. Geophys. Res.*, *96*, 9295–9308, 1991.

Eisele, F. L., D. J. Tanner, C. A. Cantrell, and J. G. Calvert, Measurements and steady state calculations of OH concentrations at Mauna Loa Observatory, *J. Geophys. Res.*, *101*, 14665–14679, 1996.

Frost, G. J., et al., Photochemical modeling of OH levels during the first aerosol characterization experiment (ACE 1), *J. Geophys. Res.*, *104*, 16041–16052, 1999.

George, L. A., T. M. Hard, and R. J. O'Brien, Measurement of free radicals OH and HO_2 in Los Angeles smog, *J. Geophys. Res.*, *104*, 11643–11655, 1999.

Goldstein, A. H., S. C. Wofsy, and C. M. Spivakovsky, Seasonal variations of nonmethane hydrocarbons in rural New England: Constraints on OH concentrations in northern midlatitudes, *J. Geophys. Res.*, *100*, 21023–21033, 1995.

Heicklen, J., Discussion of "Hydrocarbon reactivities and nitric oxide conversion" by E. R. Stephens, in C. S. Tuesday (Ed.), *Chemical Reactions in Urban Atmospheres*, Elsevier, 1971, pp. 55–59.

Impey, G. A., C. M. Mihele, P. B. Shepson, D. R. Hastie, K. G. Anlauf, and L. A. Barrie, Measurements of photolyzable halogen compounds and bromine radicals during the Polar Sunrise Experiment 1997, *J. Atmos. Chem.*, *34*, 21–37, 1999.

Jacob, D. J., *Introduction to Atmospheric Chemistry*, Princeton University Press, Princeton, NJ, 1999.

Jacob, D. J., and M. R. Hoffmann, A dynamic model for the production of H^+, NO_3^-, and SO_4^{2-} in urban fog, *J. Geophys. Res.*, *88*, 6611–6621, 1983.

Jacob, D. J., S. Sillman, J. A. Logan, and S. C. Wofsy, Least-independent-variables method for simulation of tropospheric ozone, *J. Geophys. Res.*, *94*, 8497–8509, 1989.

Jaeglé, L., et al., Observed OH and HO_2 in the upper troposphere suggest a major source from convective injection of peroxides, *Geophys. Res. Lett.*, *24*, 3181–3184, 1997.

Jaeglé, L., et al., Photochemistry of HO_x in the upper troposphere at northern midlatitudes, *J. Geophys. Res.*, *105*, 3877–3892, 2000.

Jobson, B. T., H. Niki, Y. Yokouchi, J. Bottenheim, F. Hopper, and R. Leaitch, Measurements of C_2–C_6 hydrocarbons during the Polar Sunrise 1992 Experiment: Evidence for Cl atom and Br atom chemistry, *J. Geophys. Res.*, *99*, 25355–25368, 1994.

Keene, W. C., A. A. P. Pszenny, D. J. Jacob, R. A. Duce, J. N. Galloway, J. J. Schultz-Tokos, H. Sievering, and J. F. Boatman, The geochemical cycling of reactive chlorine through the marine troposphere, *Global Biogeochem. Cycles*, *4*, 407–430, 1990.

Keene, W. L., D. J. Jacob, and S.-M. Fan, Reactive chlorine: A potential sink for dimethylsulfide and hydrocarbons in the marine boundary layer, *Atmos. Environ.*, *30*, i–iii, 1996.

Kerr, J. A., J. G. Calvert, and K. L. Demerjian, The mechanism of photochemical smog formation, *Chem. Brit.*, *8*, 252–257, 1972.

Krol, M., P. J. van Leeuwen, and J. Lelieveld, Global OH trend inferred from methylchloroform measurements, *J. Geophys. Res.*, *103*, 10697–10711, 1998.

Leighton, P. A., *Photochemistry of Air Pollution*, Academic, New York, 1961.

Levy, H., Normal atmosphere: Large radical and formaldehyde concentrations predicted, *Science*, *173*, 141–143, 1971.

Levy, H., Tropospheric budgets for methane, carbon monoxide, and related species, *J. Geophys. Res.*, *78*, 5325–5332, 1973.

Logan, J. A., M. J. Prather, S. C. Wofsy, and M. B. McElroy, Tropospheric chemistry: A global perspective, *J. Geophys. Res.*, *86*, 7210–7254, 1981.

Lovelock, J. E., Methyl chloroform in the troposphere as an indicator of OH radical abundance, *Nature*, *267*, *32*, 1977.

Madronich, S., and C. Granier, Impact of recent total ozone changes on tropospheric ozone photodissociation, hydroxyl radicals, and methane trends, *Geophys. Res. Lett.*, *19*, 465–467, 1992.

Mak, J. E., C. A. M. Brenninkmeijer, and M. R. Manning, Evidence for a missing carbon monoxide sink based on tropospheric measurements of ^{14}CO, *Geophys. Res. Lett.*, *19*, 1467–1470, 1992.

McConnell, J. C., M. B. McElroy, and S. C. Wofsy, Natural sources of atmospheric CO, *Nature*, *233*, 187–188, 1971.

McKeen, S. A., et al., Photochemical modeling of hydroxyl and its relationship to other species during the Troposheric OH Photochemistry Experiment, *J. Geophys. Res.*, *102*, 6467–6493, 1997.

Mount, G., The measurement of tropospheric OH by long-path absorption, 1. Instrumentation, *J. Geophys. Res.*, *97*, 2427–2444, 1992.

Müller, J.-F., and G. Brasseur, Sources of upper tropospheric HO_x: A three-dimensioinal study, *J. Geophys. Res.*, *104*, 1705–1715, 1999.

Neftel, A., R. C. Bales, and D. J. Jacob, H_2O_2 and HCHO in polar snow and their relation to atmospheric chemistry, in R. Delmas (Ed.), *Ice Core Studies of Global Biogeochemical Cycles*, Springer-Verlag, Berlin, 1995, pp. 249–264.

Penkett, S. A., B. M. Jones, K. A. Brice, and A. E. Eggleton, The importance of atmospheric ozone and hydrogen peroxide in oxidizing sulfur dioxide in cloud and rainwater, *Atmos. Environ.*, *13*, 123–137, 1979.

Poppe, D., J. Zimmermann, and H. P. Dorn, Field data and model calculations for the hydroxyl radical, *J. Atmos. Sci.*, *52*, 3402–3407, 1995.

Prather, M. J., and C. M. Spivakovsky, Tropospheric OH and the lifetimes of hydrochlorofluorocarbons, *J. Geophys. Res.*, *95*, 18723–18729, 1990.

Prather, M. J., and D. J. Jacob, A persistent imbalance in HO_x and NO_x photochemistry in the upper troposphere driven by deep tropical convection, *Geophys. Res. Lett.*, *24*, 3189–3192, 1997.

Prinn, R. G., R. F. Weiss, B. R. Miller, J. Huang, F. N. Alyea, D. M. Cunnold, P. J. Fraser, D. E. Hartley, and P. G. Simmonds, Atmospheric trends and lifetime of CH_3CCl_3 and global OH concentrations, *Science*, *269*, 187–192, 1995.

Pszenny, A. A. P., W. C. Keene, D. Jacob, S. Fan, J. R. Maben, M. P. Zetwo, M. Springer-Young, and J. N. Galloway, Evidence of inorganic chlorine gases other than hydrogen chloride in marine surface air, *Geophys. Res. Lett.*, *20*, 699–702, 1993.

Sigg, A., and Neftel, Evidence for a 50% increase in H_2O_2 over the past 200 years from a Greenland ice core, *Nature*, *351*, 557–559, 1991.

Singh, H. B., Atmospheric halocarbons: Evidence in favor of reduced average hydroxyl radical concentration in the troposphere, *Geophys. Res. Lett.*, *4*, 101–104, 1977.

Singh, H. B., M. Kanakidou, P. J. Crutzen, and D. J. Jacob, High concentrations and photochemical fate of oxygenated hydrocarbons in the global troposphere, *Nature*, *378*, 50–54, 1995.

Singh, H. B., and J. F. Kasting, Chlorine-hydrocarbon photochemistry in the marine troposphere and lower stratosphere, *J. Atmos. Chem.*, *7*, 261–286, 1988.

Singh, H. B., et al., Low ozone in the marine boundary layer of the tropical Pacific Ocean: Photochemical loss, chlorine atoms, and entrainment, *J. Geophys. Res.*, *101*, 1907–1917, 1996.

Spicer, C. W., E. G. Chapman, B. J. Finlayson-Pitts, R. A. Plastridge, J. M. Hubbe, J. D. Fast, and C. M. Berkowitz, First observations of Cl_2 and Br_2 in the marine troposphere, *Nature*, *394*, 353–356, 1998.

Spivakovsky, C. M., et al., Three-dimensional climatological distribution of tropospheric OH: Update and evaluation, *J. Geophys. Res.*, *105*, 8931–8980, 2000.

Staffelbach, T., A. Neftel, B., Stauffer, and D. J. Jacob, Formaldehyde in polar ice cores: A possibility to characterize the atmospheric sink of methane in the past? *Nature*, *349*, 603–605, 1991.

Thompson, A. M., The oxidizing capacity of the earth's atmosphere: Probable past and future changes, *Science*, *256*, 1157–1165, 1992.

Thompson, A. M., Measuring and modeling the tropospheric hydroxyl radical (OH), *J. Atmos. Sci.*, *52*, 3315–3327, 1995.

Toumi, R., BrO as a sink for dimethylsulfide in the marine atmosphere, *Geophys. Res. Lett.*, *21*, 117–120, 1994.

Vogt, R., P. J. Crutzen, and R. Sander, A mechanism for halogen release from sea-salt aerosol in the remote marine boundary layer, *Nature*, *383*, 327–330, 1996.

Volz, A., D. E. Kley, and R. G. Derwent, Seasonal and latitudinal variation of ^{14}CO and the tropospheric concentrations of OH radicals, *J. Geophys. Res.*, *86*, 5163–5171, 1981.

Wang, Y., and D. J. Jacob, Anthropogenic forcing on tropospheric ozone and OH since preindustrial times, *J. Geophys. Res.*, *103*, 31123–31135, 1998.

Wang, Y., D. J. Jacob, and J. A. Logan, Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry, 1. Model formulation, *J. Geophys. Res.*, *103*, 10713–10726, 1998a.

Wang, Y., J. A. Logan, and D. J. Jacob, Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry, 2. Model evaluation and global ozone budget, *J. Geophys. Res.*, *103*, 10727–10756, 1998b.

Wayne, R. P. (Ed.), The nitrate radical: Physics, chemistry, and the atmosphere, *Atmos. Environ.*, *25*, 1–203, 1991.

Weinstock, B., Carbon monoxide: Residence time in the atmosphere, *Science*, *166*, 224–225, 1969.

Weinstock, B., and H. Niki, Carbon monoxide balance in nature, *Science*, *176*, 290–292, 1972.

Wennberg, P. O., et al., HO_x, NO_x, and the production of ozone in the upper troposphere, *Science*, *279*, 49–53, 1998.

World Meteorological Association (WMO), *Scientific Assessment of ozone depletion: 1998*, WMO, Geneva, Switzerland, 1999.

CHAPTER 3

TROPOSPHERIC OZONE

JACK FISHMAN

1 INTRODUCTION

Ozone is the triatomic form of oxygen, O_3, and is generally regarded as the most important species that determines the oxidizing capacity of the troposphere. The word *ozone* comes from the Greek word *ozein*, which means "to smell." Probably the name of this gas originated from early laboratory studies when ozone was first discovered because of its distinctive acrid odor. The German scientist Christian Friedrich Schönbein is credited with ozone's discovery in 1839, while he was a professor at the University of Basel in Switzerland.

One of the goals of Schönbein's research was to show that ozone is a permanent and natural component of the atmosphere. He devised a method to measure ozone in the atmosphere that was capable of measuring very low levels simply and easily. The method used soon became known as *Schönbein paper* and involved the simple process of saturating a strip of paper with potassium iodide (KI) and then allowing it to dry. In the presence of ozone, the potassium iodide oxidized and is converted to potassium iodate (KIO_3). In the process of this conversion the paper changes color to various hues of blue. More ozone present in the atmosphere resulted in the paper becoming a deeper shade of blue. Schönbein calibrated the amount of color change into a measurement standard called *Schönbein* units, which allowed scientists to put out a new piece of Schönbein paper each day and measure the relative amount of ozone in the atmosphere.

Although the methods of measurement have been modified over the years, scientists continued to use KI to measure ozone for more than a century. One modification involved pumping ambient air through a KI solution and measuring the amount of iodide being converted to iodate since an electrical current is created in the solution

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

as the conversion takes place. The reaction took place within a matter of seconds, and the amount of electric current was easily quantifiable. This method, known as the *wet method*, was the predominant way ozone was measured until the 1960s, when other methods using newer optical technology became available and increased the accuracy of the measurements. One problem with the wet method was that other chemicals in the atmosphere interfered with the chemical reaction. The most common of these interfering trace gases is sulfur dioxide (SO_2), a pollutant that is primarily a by-product of coal combustion.

In the early part of the twentieth century, ground-based and balloon-borne measurements discovered that most of the atmosphere's ozone is located in the stratosphere with highest concentrations located between 15 and 30 km. For a long time, it was believed that tropospheric ozone originated from the stratosphere and that most of it was destroyed by contact with Earth's surface. Ozone was known to be produced by the photodissociation of molecular oxygen, O_2, a process that can only occur at wavelengths shorter than 242 nm. The atomic oxygen formed as a product of this photodissociation would then recombine with another oxygen molecule to make ozone. Because such short-wavelength radiation is present only in the stratosphere, no tropospheric ozone production is possible by this mechanism. In the 1940s, however, it became obvious that production of ozone was also taking place in the troposphere. The overall reaction mechanism was eventually identified by Arie Haagen-Smit of the California Institute of Technology located in highly polluted southern California. The smog chemistry hypothesized by Haagen-Smit was still thought to be a relatively small source on the global scale since ~90% of the ozone was located in the stratosphere, creating a ubiquitous source of tropospheric ozone as stratosphere air was transported into the troposphere. It was not until the 1970s that this viewpoint was challenged when Paul Crutzen (Crutzen, 1974) and other scientists at the time showed that consideration of "smog chemistry" in the background troposphere could produce a sizable source of tropospheric ozone and must be included in the global tropospheric ozone budget. Crutzen's pioneering work on tropospheric ozone was noted when he received the Nobel Prize for Chemistry in 1995.

2 CHEMISTRY OF TROPOSPHERIC OZONE FORMATION

Photodissociation of NO_2 by (visible) sunlight is the only significant anthropogenic source of O_3 in the troposphere

$$NO_2 + h\nu \rightarrow NO + O \qquad (\lambda < 420\ \text{nm}) \tag{1}$$

immediately followed by

$$O_2 + O + M \rightarrow O_3 + M \tag{2}$$

where the M in reaction (2) represents any nonreactive molecule that absorbs some of the excess energy of the intermediate product formed in the reaction (2).

The atmospheric oxidation of a hydrocarbon, RH, is initiated by reaction with the hydroxyl radical (OH):

$$RH + OH \rightarrow R + H_2O \tag{3}$$

where RH can be any molecule containing a hydrogen and a carbon, including methane, CH_4, or any nonmethane hydrocarbon consisting of more than one carbon atom. The product is another radical, denoted R, and water vapor. The radical quickly combines with an oxygen molecule in a three-body reaction:

$$R + O_2 + M \rightarrow RO_2 + M \tag{4}$$

to form another oxygenated radical, RO_2, called a peroxy radical. The peroxy radicals are the key for converting NO to NO_2:

$$RO_2 + NO \rightarrow RO + NO_2 \tag{5}$$

In addition, RO attaches to an oxygen molecule to form another peroxy radical:

$$RO + O_2 \rightarrow HO_2 + R'CHO \tag{6}$$

where R'CHO is an aldehyde (and noting that R' is a shorter chained carbon radical than R). The HO_2 likewise reacts with NO to form another NO_2 molecule:

$$HO_2 + NO \rightarrow OH + NO_2 \tag{7}$$

where the two NO_2 molecules photolyze and eventually produce ozone:

$$NO_2 + h\nu \rightarrow NO + O \quad \text{(2 times)} \tag{1}$$
$$O_2 + O + M \rightarrow O_3 + M \quad \text{(2 times)} \tag{2}$$

Net: $$RH + 4O_2 + 2h\nu \rightarrow R'CHO + 2H_2O + 2O_3.$$

Additional ozone molecules can also be produced through the oxidation of R'CHO. In this reaction sequence, it is important to note that the nitrogen oxide emitted as a pollutant is still available to make more ozone. If NO were not present in the atmosphere, ozone would not be formed. In fact, the presence of many nonmethane hydrocarbons, by themselves, would result in a destruction of ozone since they, or some of their daughters of the oxidation process, could react with any ozone present in the atmosphere.

On the other hand, if only nitrogen oxides and ozone were present in the atmosphere, an equilibrium would quickly be established since O_3 reacts quickly with NO:

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{8}$$

and the ratio among NO, NO_2, and O_3 is quickly established by the rates of the reactions among these species:

$$[NO]/[NO_2] = j_1/[O_3]k_8$$

where the brackets denote the concentration of a particular species, j_1 is the rate of photolysis of NO_2, and k_8 is the rate of reaction (8); the relationship among these three gases defined by this ratio is often referred to as the photostationary state and has had an important implication for understanding the formation of ozone near urban areas and subsequent strategies developed for the reduction of ozone concentrations.

3 GLOBAL DISTRIBUTION OF TROPOSPHERIC OZONE

The distribution of tropospheric ozone can be determined from the analyses of satellite data sets obtained independently from two different instruments: The Total Ozone Mapping Spectrometer (TOMS) and the Stratospheric Aerosol and Gas Experiment (SAGE). Between October 1978 and May 1993, TOMS functioned on the *Nimbus 7* satellite and provided daily maps of the distribution of total ozone. Additional TOMS were launched in 1991 (on the Russian *Meteor* satellite) and two in 1996 (see Chapter 21; "Stratospheric Ozone Observations"). The National Aeronautical and Space Administration's (NASA's) Earth Observing System (EOS) now is operational and total ozone will be measured as part of EOS. Total ozone is defined as the integrated amount of ozone between the surface and the top of the atmosphere. A unit of measure for total ozone is a quantity known as the Dobson unit (DU), where 1 DU = 2.69×10^{16} molecules O_3/cm^2. If this amount of ozone were brought down to standard atmospheric temperature and pressure, the depth of this column would be 1 mm. Thus, another common measure of column ozone is mm-atm, where a mm-atm is equivalent to 1 DU. A typical amount of total ozone found in the atmosphere is 300 DU, and approximately 90% of this ozone is located in the stratosphere.

At middle and high latitudes, the distribution of total ozone is primarily governed by the prevailing large-scale circulation patterns. These patterns can vary substantially on a daily basis, and intense gradients of total ozone have been observed with differences of 200 DU at locations less than a few thousand kilometers apart. At these higher latitudes, total ozone amounts can range between ~225 and ~500 DU. Only recently have values as low as 100 DU been observed during austral spring in conjunction with the Antarctic ozone hole.

At lower latitudes, however, the total ozone distribution patterns exhibit much smaller gradients than at middle and high latitudes. The intense gradients of as much as 200 DU found at the higher latitudes are replaced by much more subtle gradients of no more than 20 to 30 DU. Because the primary intent of the measurement of total ozone was to study the distribution of stratospheric ozone, very little research was conducted using the information provided by TOMS in the tropics. Subsequently, however, it has been shown that the variations in total ozone at low latitudes were

primarily the result of variability of ozone in the troposphere even though only ~10% of the total ozone was in the troposphere.

The use of TOMS for tropospheric studies has taken a substantive step further when data from SAGE were used to derive the amount of ozone in the stratosphere (Fishman et al., 1990). Ozone measurements from the SAGE instruments (SAGE was launched in February 1979 and operated through November 1981; SAGE II was launched in November 1984 and is still operating) provide the vertical distribution of ozone in the stratosphere. From these profiles, the amount of ozone in the stratosphere can be integrated and then subtracted from the co-located total ozone amount derived independently from the TOMS on the same day.

The distribution of the integrated amount of tropospheric ozone as a function of season is shown in Figure 1 (Fishman et al., 2002). These seasonal depictions show that there is considerably more ozone in the Northern Hemisphere than in the Southern Hemisphere, especially during the summer. During most of the seasons, distinct plumes that seem to result from pollution originating in North America, Asia, Africa, and Europe can be observed. In the three northern continents, the plumes originate over the eastern portions of each landmass and are transported by the prevailing westerly winds for several thousand kilometers. At low latitudes, the highest concentrations of pollution are off the west coast of Africa and is most pronounced during austral spring (September–November). At these latitudes, the prevailing low-level winds are trade winds (easterlies), which would carry the emissions from central and western Africa to the eastern tropical South Atlantic Ocean. The prevailing upper level winds are westerlies, so any ozone that gets to altitudes of ~5 km or higher are transported long distances to the east. Evidence of the long-range transport of emissions from biomass burning in Africa and South America to Australia is evident in long-term Australian data sets of not only ozone but also carbon monoxide and elemental carbon, two other products of widespread burning.

4 TROPOSPHERIC OZONE TRENDS IN NONURBAN TROPOSPHERE

The global distribution of tropospheric ozone shown in Figure 1 illustrates its wide range (more than a factor of 3) of abundance. Therefore, unlike trace gases such as chlorofluorocarbons, nitrous oxide, or carbon dioxide, which exhibit very small spatial gradients, an assessment of the *global* rate of increase of tropospheric ozone is difficult to determine from measurements at only a few locations. Outside of urban areas, only a few stations around the world have continuous long-term measurements of tropospheric ozone. Among these stations are the ones set up by the U.S. National Oceanographic and Atmospheric Administration (NOAA), which has maintained a carefully calibrated monitoring program at a number of stations around the world since the early 1970s (Oltmans and Levy, 1994). The monthly mean concentrations from Barrow and Mauna Loa are shown on the left side of Figure 2*a*. The linear least-squares fit illustrating the trend between 1973 and 1992 for these two data sets is also plotted on these figures. Even though both of these stations show a significant increase over this period, the measurements at Barrow

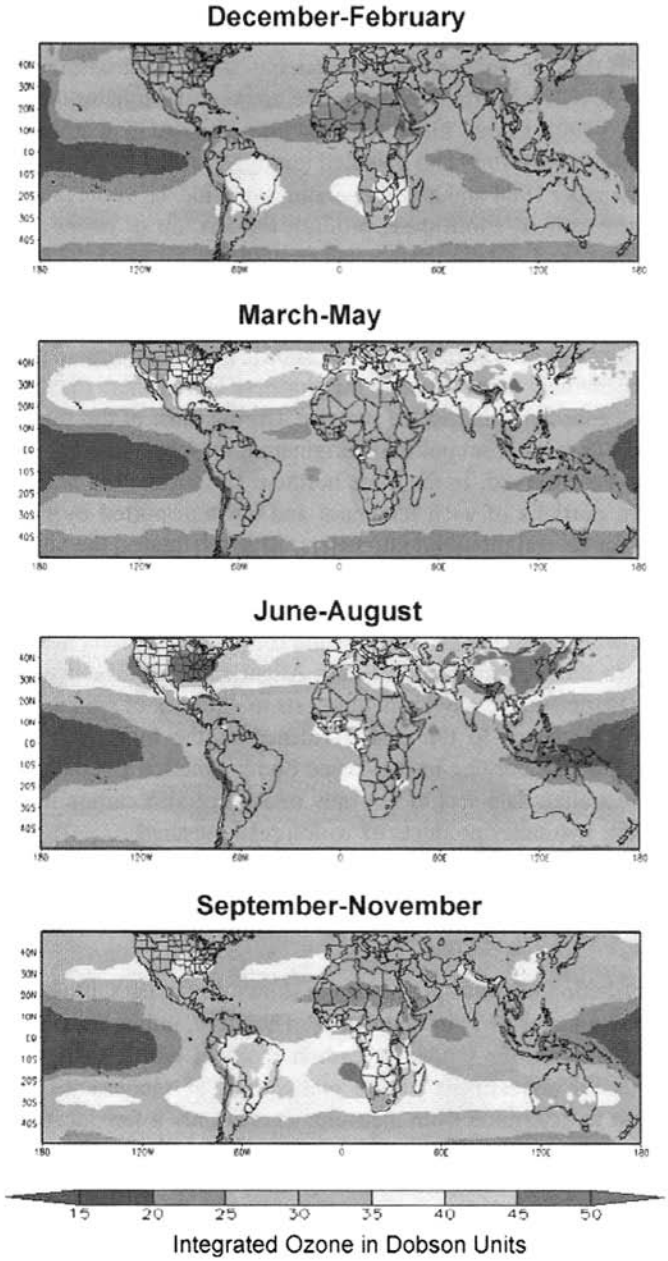


Figure 1 (see color insert) Climatological distribution of tropospheric ozone derived from satellite measurements between 1979 and 2000 (from Fishman et al., 2002). Units of contours and Dobson Units (DU). Regions greater than 40 DU have been shaded. See ftp site for color image.

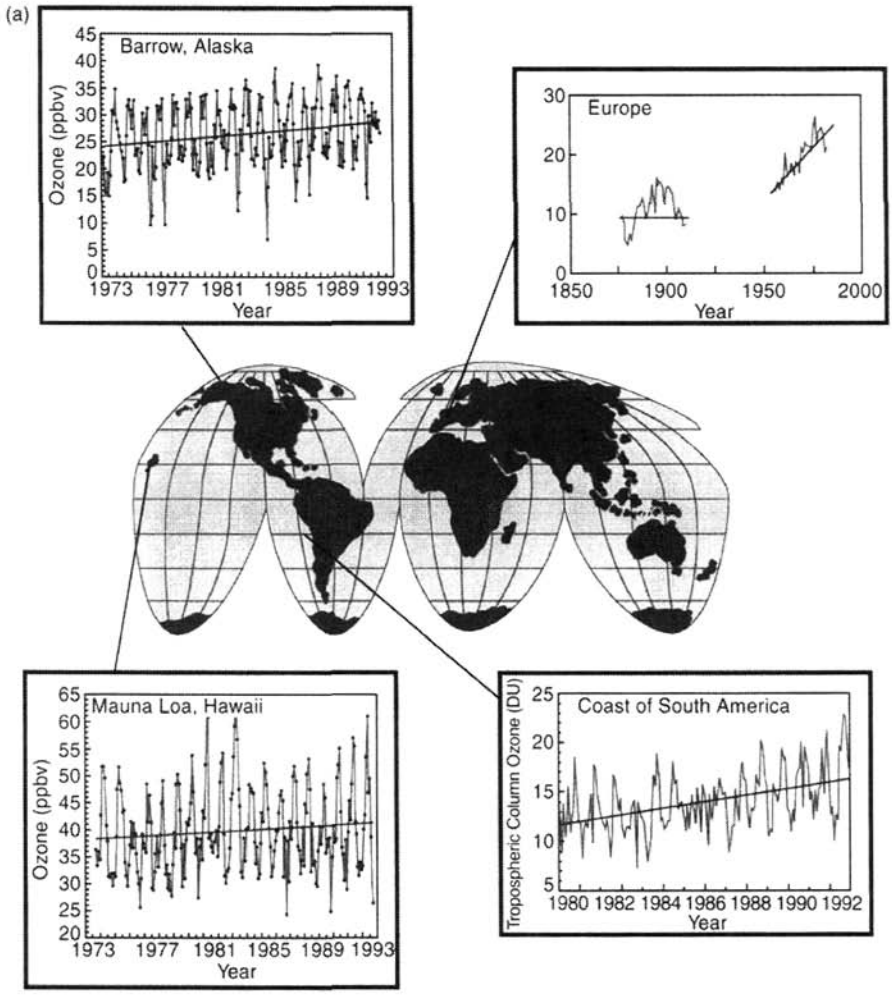


Figure 2 (*a*) (Upper left): monthly mean surface ozone at Barrow and the linear trend for the entire data record. (Lower left): monthly mean surface ozone at Mauna Loa with the linear trend. (Upper right) annual mean ozone concentrations at Montsouris Observatory outside Paris (1876–1910) and Arkona, East Germany (1956–1984). The average ozone concentration at the beginning of the twentieth century near Paris was less than 10 ppb whereas in 1985 the typical ground-level concentrations in central Europe is approaching 30 ppb, implying an increase of about 200% during the century (from Volz and Kley, 1988). (Lower right): ozone trend off coast of South America determined from analysis of satellite measurements of total ozone (from Jiang and Yung, 1996). See ftp site for color image.

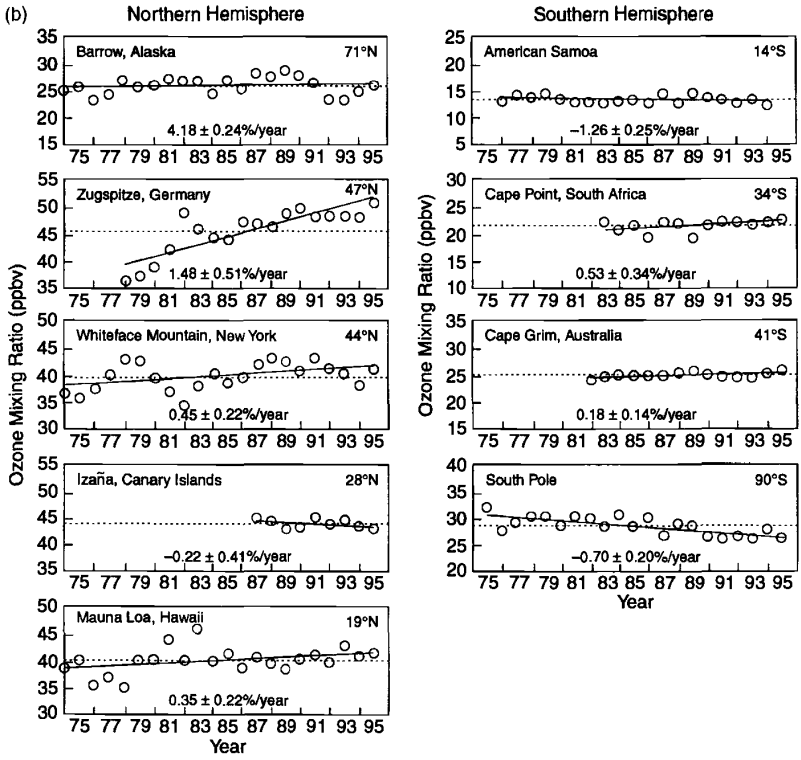


Figure 2 (*b*) Annual average ozone mixing ratios (ppbv) for surface ozone measuring sites. The dashed line is the long-term average. The solid line is the linear least-squares fit to the plotted values. The linear trend and 95% confidence in percent per year is given with each location (from Oltmans et al., 1998).

show that the long-term trend has a strong seasonal dependence; the increase during the summer is 1.73% per year whereas there is almost no trend (−0.07% per year) during the winter. Figure 2*b* summarizes a number of long-term measurements from the NOAA network as well as a few other stations where comparable data exists in the background atmosphere. Curiously, some stations such as American Samoa near the equator show a slight negative trend, whereas a significant negative trend exists at South Pole. The reason for these trend differences at these remote sites is not clear and is currently being studied.

Modern studies have reexamined the Schönbein paper ozone measurements from the late nineteenth century and early twentieth century to determine tropospheric ozone trends over longer time periods. These studies have carefully examined calibration procedures used last century and have determined that a significant increase in tropospheric ozone has occurred over the past century.

More than three decades of measurements using Schönbein's technique were obtained at the Montsouris Observatory outside Paris. The instrument used at this meteorological station was recalibrated and the observations were converted to standard units of measurement consistent with modern measurements. The results from this data set are compared with modern observations obtained in Germany and depicted in the upper right panel of Figure 2*a*. This and other analyses strongly suggest that ozone at the surface has risen from ~10 ppbv to more than 30 ppbv in nonurban Europe and the eastern United States. Although ozone at the surface has likely increased significantly on the time scales of years and decades since the inception of the industrial era, tropospheric measurements above the surface are extremely scarce and difficult to interpret because of the different methods of measurement used since the 1960s. Most of the measurements are from ozonesondes (an ozone sensor placed on a balloon), but several types of sensors have been used and each type is susceptible to interference from other trace gases in the atmosphere. Despite the uncertainty in the measurements, it is generally believed that ozone has increased throughout the entire troposphere since the 1960s, when ozonesonde measurements started on a fairly regular basis.

5 GLOBAL TROPOSPHERIC OZONE BUDGET

The components of the global tropospheric ozone budget can be broken into four general categories: transport from the stratosphere, destruction at Earth's surface, photochemical destruction, and in situ photochemical production. The primary mechanism by which ozone is transported from the stratosphere into the troposphere is through meteorological events referred to as stratospheric intrusions. These events occur in conjunction with the movement of air associated with rapid changes in the intensity and position of the jet stream, the fast-moving westerly river of air that often delineates the position of strong frontal boundaries at middle latitudes. Under these conditions, the tropopause (i.e., the boundary between the troposphere and the stratosphere) often becomes contorted and its position becomes difficult to define and often takes on a "folded" depiction (see Chapter 1, "Overview: Atmospheric Chemistry). Because of this, stratospheric intrusions are also synonymous with tropopause folding events.

The topic of stratosphere–troposphere exchange was an intense research area in the 1960s and early 1970s because of the concern of transport of radioactive debris created by atmospheric nuclear bomb testing from the stratosphere into the lower atmosphere and eventually its deposition to plants, animals and human populations. During this time, the North American Ozonesonde Network was established for the primary purpose of understanding how stratospheric air was transported into the troposphere. From these data, it is generally thought that ~10% of the stratosphere is exchanged annually with the troposphere. From these estimates, the global source of tropospheric ozone from the stratosphere, which was assumed the primary *natural* source of tropospheric ozone could be computed (e.g., Danielsen and Mohnen, 1977).

The other primary component of the global budget of tropospheric ozone is its sink, or how it is destroyed once it is in the troposphere. The early measurements of

ozone's vertical distribution always showed that lowest concentrations were near Earth's surface, implying a sink for ozone as it came in contact with the ground. These measurements generally showed much sharper vertical gradients over land and vegetated surfaces than over water and ice surfaces. Thus, one way to determine this deposition sink globally was to make a series of field measurements over a representative sample of surfaces and extrapolate these measurements to the rest of the world. Using this methodology, the globally averaged destruction rate of tropospheric ozone generally converged to a value near 8 to 10×10^{10} molecules O_3/cm^2 s. The accuracy of these estimates was claimed to be ~30%. These calculations were consistent with the few attempts to extrapolate the global input from the stratosphere resulting from stratosphere–troposphere exchange studies, which indicated that a global average of $\sim 8 \times 10^{10}$ molecules O_3/cm^2 s came from the stratosphere. Thus, up until the early 1970s, it was generally believed that the tropospheric ozone budget was balanced by the natural input from the stratosphere and the destruction at Earth's surface (Fabian and Junge, 1970). The potential impact of local-scale photochemical generation (as was known at the time for areas such as southern California) was believed to be insignificant.

A series of studies published shortly thereafter challenged this assumption and proposed that a natural source of tropospheric ozone of comparable magnitude to that of input from the stratosphere existed in the background atmosphere as a result of methane oxidation. For the first time, the paradigm of the tropospheric ozone budget was challenged resulting in a lively debate in the scientific literature in the middle and late 1970s (Chameides and Walker, 1973; Fabian, 1973; Fishman and Crutzen, 1978). These theoretical studies primarily concentrated on the generation of ozone from the oxidation of methane and carbon monoxide, the two most abundant trace gases that could lead to the photochemical formation of tropospheric ozone.

Another important component of the tropospheric ozone budget is its photochemical destruction. As ozone enters from the stratosphere, for example, it is photolyzed at shorter wavelengths to produce an excited state of atomic oxygen, $O(^1D)$, rather than its ground state, $O(^3P)$:

$$O_3 + h\nu \rightarrow O(^1D) + O_2 \qquad (\lambda < 320 \text{ nm}) \tag{9}$$

Once $O(^1D)$ is formed, it can react with water vapor to generate OH:

$$O(^1D) + H_2O \rightarrow 2OH \tag{10}$$

In turn, OH can react with ozone to form the hydroperoxy radical, which can set up a catalytic cycle of ozone destruction, analogous to what happens in the stratosphere:

$$OH + O_3 \rightarrow HO_2 + O_2 \tag{11}$$

followed by $$\underline{HO_2 + O_3 \rightarrow OH + 2O_2} \tag{12}$$

$$2O_3 \rightarrow O_2 \quad \text{(net reaction sequence)}$$

Reaction (10) is the primary source of OH in the troposphere, and subsequent reactions with OH are the primary means by which most chemicals released to the atmosphere are oxidized and eventually removed. Whereas photochemistry was first proposed as an important photochemical source of tropospheric ozone in the studies written in the early 1970s, it is important to also note that photochemistry is also the dominant sink and is the primary reason that ozone concentrations are generally very low in the tropical troposphere where both water vapor and incoming solar flux are highest. The key to whether photochemistry is a net source or a net sink for tropospheric ozone is most dependent on how much NO is present.

6 CURRENT UNDERSTANDING OF TROPOSPHERIC OZONE BUDGET

The global distribution of tropospheric ozone presented earlier in this chapter illustrates its heterogeneity and underscores the difficulty of quantifying a global budget using the simplistic assumptions about its vertical distribution that had been employed when budgets neglecting photochemical processes were formulated. It is clear from the depiction in Figure 1 that local-scale photochemical generation of ozone has had a considerable impact on the global distribution as evidenced by the dominant plumes originating over North America, Europe, Asia, and Africa. A proper calculation of the tropospheric ozone budget must quantify these local- and regional-scale processes that feed into the global budget. Studies investigating photochemical processes from industrial emissions of volatile organic compounds and nitrogen oxides on scales of ~1000 km showed that the ozone generated on these scales should at least be comparable to the amount generated in the background through methane and carbon monoxide oxidation. In addition, the data now indicate that large quantities of ozone are generated in the tropics as emissions from widespread vegetation burning are oxidized efficiently in the intense tropical sunshine. Furthermore, some recent analyses of ozonesonde data have concluded that very little (perhaps as small as 5%) ozone near the ground had originated in the stratosphere and only ~25% of the ozone observed at 300 mbar had originated in the stratosphere. This analysis agrees with more recent estimates of stratosphere–troposphere mass exchange suggesting that the amount of ozone from the stratosphere is likely only ~30% of the amount determined from the earlier estimates determined in the 1970s.

Calculations from a general circulation model, which includes a complete set of photochemical reactions, have been used to evaluate the tropospheric ozone budget (Wang et al., 1998). The results from these model calculations are shown in the four seasonal panels in Figure 3. These calculations show how the chemical terms are both considerably larger than the input from the stratosphere and the amount of destruction at the ground. In addition, the amount of ozone produced photochemically is generally greater than the amount destroyed. The largest amount of production is at northern middle latitudes in July. The Southern Hemisphere is also a sizable source in both July and October, when biomass burning is most prevalent in the southern tropics and subtropics.

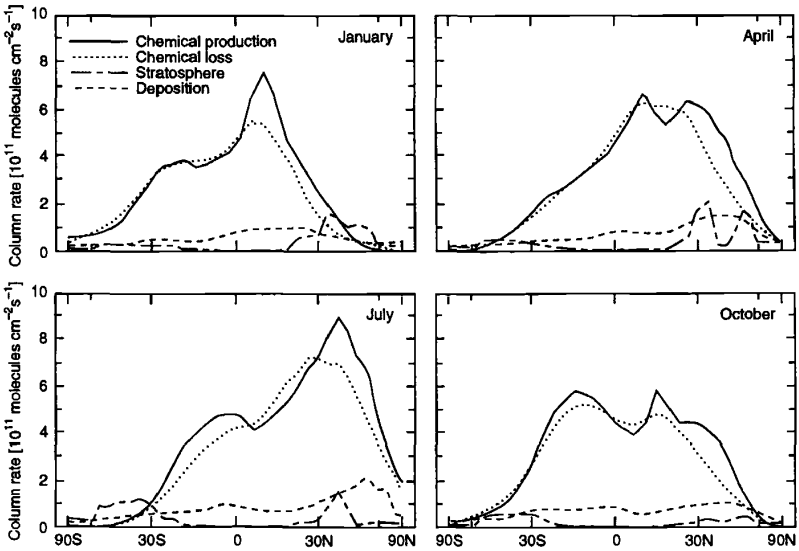


Figure 3 Zonally averaged column budget for tropospheric ozone in different seasons including term from in situ photochemical production and loss, transport from the stratosphere, and deposition. The abscissa scale is linear in sine of latitude (from Wang et al., 1998).

These studies, as well as the documented increase in tropospheric ozone over time scales of decades provide fairly strong evidence that its distribution has changed significantly over the last century and that a large fraction of the tropospheric ozone budget is now likely controlled by anthropogenic pollution from both industrialized and tropical regions of the world. Studies are currently underway to provide more quantitative information, and our understanding of tropospheric ozone will greatly improve as more data are analyzed and more sophisticated global models are developed to study the problem.

REFERENCES

Chameides, W. L., and J. C. G. Walker, A photochemical theory of tropospheric ozone, *J. Geophys. Res.*, *78*, 8751–8760, 1973.

Crutzen, P. J., Photochemical reactions initiated by and influencing ozone in unpolluted tropospheric air, *Tellus*, *26*, 47–57, 1974.

Danielsen, E. F., and V. A. Mohnen, Project dustorm report: Ozone transport, in situ measurements and meteorological analyses of tropopause folding, *J. Geophys. Res.*, *82*, 5867–5877, 1977.

Fabian, P., A theoretical investigation of tropospheric ozone and stratospheric-tropospheric exchange processes, *Pure Appl. Geophys.*, *106–108*, 1044–1057, 1973.

Fabian, P., and C. E. Junge, Global rate of ozone distribution at the earth's surface, *Arch. Meteor. Geophys. Biokl. Ser. A.*, *19*, 161–172, 1970.

Fishman, J., and P. J. Crutzen, The origin of ozone in the troposphere, *Nature*, *274*, 855–858, 1978.

Fishman, J., C. E. Watson, J. C. Larsen, and J. A. Logan, Distribution of tropospheric ozone determined from satellite data, *J. Geophys. Res*, *95*, 3599–3617, 1990.

Fishman, J., A. E. Balok, and F. M. Vukovich, Observing tropospheric trace gases from space: recent advances and future capabilities, *Adv. Space Res.* **29**, 1625–1630, 2002.

Jiang, Y., and Y. L. Yung, Concentrations of tropospheric ozone from 1979 to 1992 over tropical Pacific South America from TOMS data, *Science*, **272**, 745–748, 1996.

Oltmans, S. J., and H. Levy II, Surface ozone measurements from a global network, *Atmos. Environ.*, *28*, 9–24, 1994.

Oltmans, S. J., et al., Trends of tropospheric ozone in the troposphere, *Geophys. Res. Lett.*, *25*, 139–142, 1998.

Volz, A., and D. Kley, Evaluation of the Montsouris series of ozone measurements made in the nineteenth century, *Nature*, *252*, 240–242, 1988.

Wang, Y., D. J. Jacob, and J. A. Logan, Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry, 3. Origin of tropospheric ozone and effects of nonmethane hydrocarbons, *J. Geophys. Res.*, *103*, 10757–10767, 1998.

CHAPTER 4

NITROGEN OXIDES AND OTHER REACTIVE NITROGEN SPECIES

J. H. CRAWFORD, J. D. BRADSHAW, D. D. DAVIS, AND S. C. LIU

1 INTRODUCTION

Nitrogen is most abundant in the atmosphere in its molecular form, N_2, which comprises 78% of Earth's atmosphere. Although virtually inert and of no direct consequence to tropospheric chemistry, this vast reservoir of atmospheric nitrogen enables the existence of trace levels of nitrogen oxides that play a number of critical roles in the chemistry of the atmosphere. Nitrogen oxides, commonly referred to as NO_x, are defined by the sum of the chemical species NO and NO_2. These two atmospheric constituents are grouped for convenience due to their fast photochemical cycling, which brings them into equilibrium generally within a few minutes. The greater family of reactive nitrogen, conventionally denoted by the term NO_y, consists of NO_x as well as a suite of other compounds including NO_3, $N2O_5$, HNO_3, HONO, HO_2NO_2, peroxyacetylnitrate (PAN), and a wide array of other organic nitrogen-containing species. These compounds play important roles in the removal of reactive nitrogen from the atmosphere as well as the transport of reactive nitrogen from source regions to remote areas.

Tropospheric chemical cycles involving NO_x are of fundamental importance to understanding several key atmospheric issues. For instance, the tropospheric ozone abundance is largely regulated by catalytic photochemical cycles involving NO_x, CO, and hydrocarbons that produce ozone (see Chapter 3). NO_x often represents the rate-limiting precursor for ozone production, especially throughout the remote atmosphere. This is due to its short lifetime relative to other precursors. On a regional scale, the role of NO_x in creating high concentrations of ozone detrimental to human health is a major air quality issue in many urban areas. On a global scale, the impact

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

of NO_x on ozone represents an important factor in determining the oxidizing capacity of the atmosphere. NO_x further impacts atmospheric oxidation rates by regulating OH concentrations, especially at high altitudes and latitudes. Since the primary mechanism for removing many pollutant gases from the atmosphere is reaction with OH, NO_x is important to the atmosphere's ability to cleanse itself. This in turn relates to the issue of climate change regarding removal of greenhouse gases such as CH_4. Another important link to the issue of climate change involves the impact of NO_x on ozone production in the upper troposphere where it is most effective as a greenhouse gas.

The tropospheric distribution of NO_x is complicated by a combination of diverse sources. Natural as well as anthropogenic sources exist both at the surface (e.g., soil emissions, biomass burning, and fossil fuel combustion) and in the free troposphere (e.g., lightning, aircraft, and stratosphere–troposphere exchange). Regeneration of NO_x through chemical recycling of various NO_y species represents a secondary source of NO_x in the troposphere. Also, the atmospheric lifetime of NO_x ranges from hours to days depending predominantly on altitude. As a result, NO_x mixing ratios vary from a few parts per trillion in some remote regions to several parts per billion in highly polluted conditions. Given the high variability of NO_x and its importance to several key atmospheric issues, the global NO_x distribution represents a pivotal subject in efforts to fully understand the current state of our atmosphere as well as its future evolution.

2 CHEMICAL TRANSFORMATIONS AND SPECIATION OF REACTIVE NITROGEN

Almost all reactive nitrogen is introduced into the atmosphere as NO, but within minutes, NO reaches equilibrium with NO_2. This NO_x is subsequently transformed into other NO_y species that can be removed, transported, or recycled back to NO_x. A general outline of these transformations is represented in Figure 1, which accompanies the following discussion of the behavior and importance of various NO_y species.

1. *NO_x ($NO + NO_2$).* During the day, NO and NO_2 experience rapid interconversion via the following simple reaction scheme:

$$NO + O_3 \rightarrow NO_2 + O_2 \quad (1)$$

$$NO_2 + h\nu \rightarrow NO + O \quad (2)$$

$$O + O_2 \rightarrow O_3 \quad (3)$$

Net: no change

This reaction sequence is a null cycle that serves no purpose photochemically other than to partition NO_x into NO and NO_2. NO may also be converted to NO_2 by hydroperoxy radicals (HO_2) that result from the oxidation of CO as well as organic

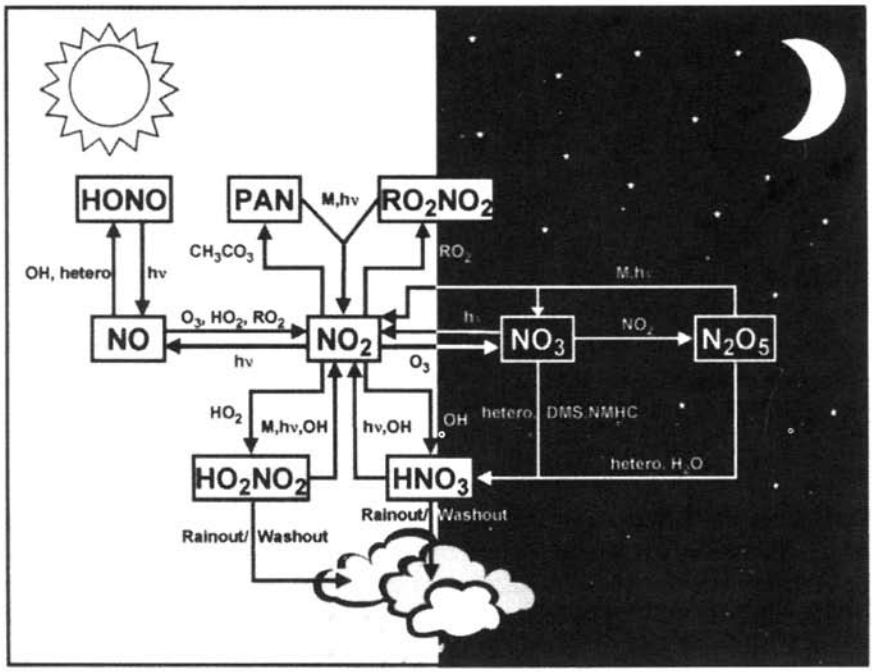


Figure 1 General schematic of reactive nitrogen chemistry in the troposphere arranged to emphasize dominant pathways in the presence of sunlight and in darkness.

peroxy radicals (RO_2, where R denotes a CH_3 or higher organic grouping) that result from the oxidation of hydrocarbons. This cycle of NO–NO_2 interconversion has impacts outside the NO_2–NO system.

$$\mathrm{NO} + \mathrm{HO_2} \rightarrow \mathrm{NO_2} + \mathrm{OH} \quad (4)$$

$$\mathrm{NO_2} + h\nu \rightarrow \mathrm{NO} + \mathrm{O} \quad (2)$$

$$\mathrm{O} + \mathrm{O_2} \rightarrow \mathrm{O_3} \quad (3)$$

$$\text{Net:} \quad \mathrm{HO_2} + \mathrm{O_2} \rightarrow \mathrm{OH} + \mathrm{O_3}$$

Here, two key impacts of NO_x interconversion are the formation of O_3 and the regeneration of OH from HO_2.

The conversion of NO_x into other longer-lived NO_y reservoir species is accomplished almost exclusively through reactions involving NO_2 (see Figure 1). Thus, the lifetime of NO_x in the atmosphere relies in part on the partitioning of NO_x between its constituents, NO and NO_2. As shown in Figure 2, the fraction of NO_x existing in the form of NO_2 changes dramatically with altitude. At the surface, NO_x tends to be predominantly in the form of NO_2 since reaction (1) proceeds at a faster rate than reaction (2). Here, NO_x lifetimes are typically one day or less. Reaction (1), however, has a strong temperature dependence and is about 5 times slower at the cold tempera-

tures of the upper troposphere, thus shifting the NO_x equilibrium in favor of NO. To a lesser degree, the increase in reaction (2) with altitude (~50%) also contributes to an NO_x partitioning that favors NO at high altitude. As NO_2 becomes a smaller fraction of NO_x with increasing altitude, the lifetime of NO_x lengthens. In the upper troposphere, NO_x lifetimes can be a few days to a week. The longer lifetime of NO_x at high altitude tends to enhance its per-molecule efficiency in the production of ozone since NO can be cycled through reaction (4) more times before being lost. Although efficiency is increased at high altitude, the ozone production rate per molecule of NO_x is slower owing to the lower abundance of HO_2, which generally decreases with altitude.

2. *NO_3*. The nitrate radical, NO_3, photolyzes within a few seconds in sunlight; thus, it is of negligible importance to the daytime photochemistry of the atmosphere. NO_3 is formed by the reaction of NO_2 with O_3.

$$NO_2 + O_3 \rightarrow NO_3 \tag{5}$$

Overnight at the surface, a significant fraction of NO_2 may be converted by this reaction. The strong temperature dependence of reaction (5), however, slows conversion rates by an order of magnitude for the upper free troposphere, where only a small fraction of NO_2 may be converted overnight. Given its concentration and high

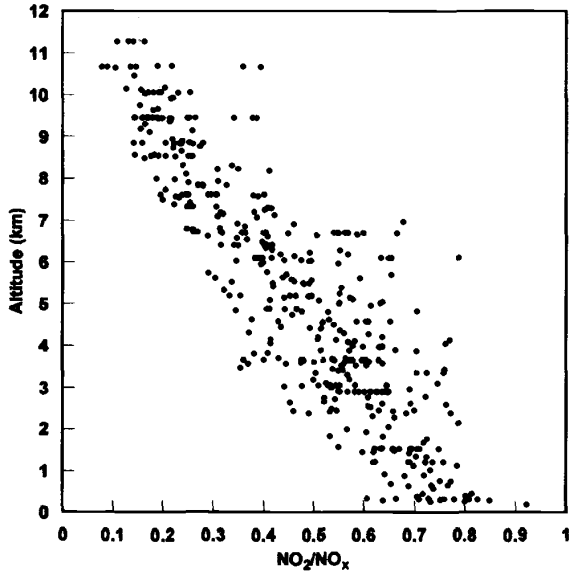


Figure 2 Fraction of NO_x in the form of NO_2 as a function of altitude. Data based on concurrent measurements of NO and NO_2 conducted during NASA's PEM–Tropics A field campaign (Bradshaw et al., 1999).

reactivity, NO_3 can be competitive with OH as an oxidant of dimethylsulfide (DMS) in the marine boundary layer of coastal regions. In general, however, marine NO_x levels are insufficient to support more than a minor role for NO_3 in DMS oxidation. In the continental boundary layer, NO_3 can be important in the oxidation of unsaturated hydrocarbons, e.g., olefins and biogenic hydrocarbons such as isoprene.

3. *N_2O_5.* N_2O_5 is formed during periods of darkness by the reaction of NO_3 and NO_2.

$$NO_3 + NO_2 + M \rightarrow N_2O_5 + M \tag{6}$$

Here, M represents an inert third body, typically N_2 or O_2. N_2O_5 is a thermally labile species; therefore, the equilibrium represented by reaction (6) favors NO_3 near the surface. Larger concentrations of N_2O_5 exist at high altitude where it is favored by colder temperatures, although its concentration is still limited by the slowdown in reaction (5). N_2O_5 also has a lifetime due to photolysis of several hours, but this is not sufficient to prevent significant daytime concentrations in the upper troposphere. N_2O_5 represents a loss of NO_x through its heterogeneous conversion to nitric acid (HNO_3) on aerosol surfaces.

4. *HNO_3.* Nitric acid is believed to be the major reservoir species for NO_x. It is formed primarily by the reaction of NO_2 and OH.

$$NO_2 + OH + M \rightarrow HNO_3 \tag{7}$$

HNO_3 can also result from the heterogeneous reaction of N_2O_5 on aerosols or the reaction of NO_3 with certain hydrocarbons. HNO_3 is efficiently removed from the atmosphere through dry deposition and rainout processes. HNO_3 may also be recycled back into NO_x through either reaction with OH or photolysis with a lifetime of a few weeks. At low altitude these two processes are slow compared to wet removal, but they can be important at high altitude where wet removal is less frequent.

5. *HONO.* Nitrous acid is formed by the gas-phase reaction of NO and OH. It also photolyzes within minutes and thus is a negligibly small component of NO_y. Although the details are not fully understood, evidence exists for nighttime formation of HONO, possibly through heterogeneous processes, based on elevated nighttime observations of HONO. Substantial buildups of HONO may take place in special environments such as overnight in polluted air rich in NO_x or in polar regions with extended periods of darkness. For these special conditions, HONO may for a short time be the dominant source of OH through its rapid photolysis during sunrise periods.

6. *HO_2NO_2.* Pernitric acid is a thermally labile species resulting from the reaction of HO_2 and NO_2. As with N_2O_5, it is favored at the cold temperatures of the upper troposphere. In the upper troposphere, HO_2NO_2 concentrations are limited by reaction with OH and photolysis resulting in a lifetime of only a couple of days, but

concentrations should approach that of NO_x. HO_2NO_2 may also return to NO_x through thermal decomposition in descending air masses.

7. *PAN.* Peroxyacetylnitrate is the most common organic nitrogen-containing species resulting from the reaction of NO_2 with the CH_3CO_3 radical. The CH_3CO_3 radical results from oxidation of a wide range of hydrocarbons, but at high altitude oxidation of acetone appears to be predominantly responsible. While loss in the lower troposphere is dominated by thermal decomposition, loss in the upper troposphere occurs through photolysis with a lifetime of 1 to 2 months. At high altitude, PAN is thermally stable and serves as an effective reservoir for global-scale transport of NO_y. In remote regions, thermal decomposition of PAN in descending air masses can be a dominant source of NO_x. Somewhat analogous to HO_2NO_2, organic peroxy radicals (RO_2) can react with NO_2 to form organic species (RO_2NO_2) with properties similar to those of PAN.

8. CH_3ONO_2. Although not depicted in Figure 1, methyl nitrate represents the most common member of a family of alkyl nitrates that can result from NO_x in the presence of hydrocarbon oxidation. There is also evidence that these species are emitted from the ocean in small amounts. Loss is primarily through photolysis to yield NO_x.

3 SOURCES OF REACTIVE NITROGEN

Unlike most trace species that are emitted only at the surface, NO_x sources exist both at the surface and in the free troposphere. Surface sources include both natural and anthropogenic sources; e.g., soil/microbial emission, fossil fuel combustion, and biomass burning. In the free troposphere, NO_x sources include lightning, aircraft emissions, and stratosphere–troposphere exchange. Table 1 gives estimated source strengths and uncertainties for each of these sources. While there are still substantial uncertainties in these sources, the global NO_x source strength is clearly dominated by surface sources with anthropogenic use of fossil fuels having the greatest contribution. The smaller sources in the free troposphere, however, cannot be trivialized since they are localized in a region of the atmosphere where NO_x lifetimes are maximized. Thus, their proportional impact is greater than their absolute source strengths would imply.

Fossil Fuel Combustion

NO is formed by high-temperature chemical processes during combustion of fossil fuels, both from nitrogen present in fuel and from the oxidation of atmospheric N_2 in the presence of O_2. The distribution for this source is heavily weighted toward the Northern Hemisphere where most of the industrialized world resides. Detailed inventories are available for Canada, the United States, and western Europe describing the spatial patterns of NO_x emissions from fossil fuel combustion and industrial processes [Wagner et al., 1986; Environmental Protection Agency (EPA), 1986;

TABLE 1 Sources of Tropospheric NO_x

Source	Estimated Magnitude and Uncertainty (Tg N/yr)	Principle Location of Emissions
Fossil fuels	22 (13–31)	Midlatitude continental surface (30–60°N)
Biomass burning	7.9 (3–15)	Tropical continental surface
Soil emissions	7.0 (4–12)	Nonpolar continental surface
Lightning	5.0 (2–20)	Tropical/subtropical continental troposphere
Aircraft	0.56 (0.45–1)	NH upper troposphere (30–60°N, 8–13 km)
Strat–Trop exchange	0.64 (0.4–1)	Midlatitude tropopause
Oceans	0.5 (0–1)	Tropical/subtropical upwelling regions
NH_3 oxidation	0.6 (0.3–3)	Free troposphere over industrial regions

Source: Adapted from Lee et al. (1997) and Bradshaw et al. (2000).

Lübkert and Zierock, 1989]. Almost one-half (44%) of the NO_x emissions in the United States are from transportation, 33% from power plants, and 16% from industrial combustion. Similar inventories have been reported for western Europe (Lübkert and DeTilly, 1989) and Asia (Akimoto and Narstu, 1994). Based on 1985 data, approximately 84% of total emissions are accounted for by emissions from North America (28%), Europe (31%), and Asia (31%).

NO_x emissions due to maritime shipping have been estimated to contribute as much as 3 Tg N/yr to the global NO_x budget (Corbett et al., 1999) with half of these emissions occurring in the North Atlantic. While small compared to the overall fossil fuel contribution of 22 Tg N/yr, NO_x emissions from seagoing vessels could prove important over the open ocean in and around shipping lanes far removed from major continental sources (Lawrence and Crutzen, 1999).

Biomass Burning

Biomass burning in tropical and subtropical regions is a significant source of NO_x as well as other chemically and radiatively important species such as CO_2, CO, CH_4, NMHC, N_2O, and aerosols. Recent estimates of NO_x emissions are based on emission factors from laboratory as well as field measurements of burning vegetation coupled with biomass inventories, land-use data, estimates of tropical deforestation, and occurrence of wild fires. Burning in the tropical latitudes is estimated to account for approximately 87% of the global total with Africa, South America, and Asia accounting for approximately 42, 23, and 28%, respectively. Thus far, however, laboratory mass-balance experiments can only account for approximately 30 to 50% of the fuel nitrogen that is released from this burning. Much of the missing

nitrogen is thought to be in the form of molecular nitrogen with the remainder possibly representing mineralized ash (10%) and high-molecular-weight compounds containing substantial amounts of nitrogen (Lobert et al., 1991; Yokelson et al., 1996). It is unclear whether these latter compounds might act as a source of NO_y or NO_x to the remote troposphere. Of the known fixed nitrogen compounds, NO_x is the dominant species (54%), having an estimated emission factor of approximately 2.1 g N/kg C (carbon fuel). This is significantly smaller than the emission factors reported for higher temperature fossil fuel combustion sources. Reduced nitrogen compounds such as NH_3 (emission factors of approximately 1.3 g N/kg C) comprise the remaining 46%.

Soil Emissions

Observations have shown that the dominant reactive odd-nitrogen emission from soils to the atmosphere is NO, with lesser emissions of NO_2 and HONO. Studies indicate that a wide range of factors influence the net soil emission of NO_x to the atmosphere. These include climate (through temperature and rainfall), plant growth and decay, the clearing of forests, biomass burning, and fertilization. The three most important variables influencing NO_x emissions are soil temperature, soil moisture content, and soil vegetation cover. NO emission rates have been found to vary almost exponentially with soil temperature, whereas a more linear relationship has been observed with respect to soil nitrate levels. The dependence on soil moisture content appears to be a complex one. Below approximately 15% soil moisture, microbial activity has been found to be primarily water limited and strongly favors nitrification, whereas at higher moisture contents denitrification eventually becomes predominant and NO emissions decrease rapidly. Order of magnitude differences in emission rates also occur between heavily fertilized soils, grasslands, and forested ecosystems (Williams et al., 1992). Large increases in the rates of soil emission have been observed after rain events following long periods of drought and in areas where biomass burning had recently occurred (Neff et al., 1995). Canopy cover has also been shown to be a key factor controlling the net flux of NO_x into the atmosphere, particularly, tropical rain forest canopies, which have been shown to be an effective sink for NO_2 (Jacob and Wofsy, 1990). Agriculture and grass lands account for the bulk of net emissions (41 and 35%, respectively) (Yienger and Levy, 1995). Future changes in soil emissions are expected to be linked to increased use of nitrogen fertilizers and agricultural production.

Lightning

Of all NO_x sources, improving our understanding of lightning is most critical since it has one of the largest uncertainties and represents what appears to be the dominant source of NO_x in the free troposphere. Lightning NO is generated by recombination reactions that occur as this 20,000 K or hotter, 10-MPa pressurized plasma supersonically expands and cools. Quantifying NO_x production from such events is still quite problematic.

Much of the current debate stems directly from estimating the amount of NO_x produced by a "typical" lightning flash. This process has involved combining estimates of the average energy deposited per lightning flash with evaluations of the NO_x produced per joule of energy released and the average global frequency of lightning. Using this approach, the global estimate of 2 Tg N/yr by Kumar et al. (1995) lies at the low end of a clustering of similarly derived estimates based on very low yields of NO_x per lightning flash ($\sim 3.6 \times 10^{25}$ NO_x molecules/flash; e.g., Lawrence et al. (1995) and references therein). At the high end, Liaw et al. (1990) have argued for a source strength as large as 200 Tg N/yr based on a correspondingly larger value for NO_x yield per lightning flash, but this estimate appears unreasonably high based on nitrate deposition records. Many of these treatments have relied on the application of scaling or normalization factors to other investigators' results that may not be valid, and frequently the various forms of lightning have been treated as if they were one type only.

Combining more refined values for the production of NO_x per joule of energy (Goldenbaum and Dickerson, 1993) with updated lightning flash energy values results in a per-flash NO_x yield for negative CG (cloud to ground) lightning of 1 to 2×10^{26} NO/flash. By contrast, for positive CG and IC (intracloud) lightning, the yield is approximately 5×10^{26} and 0.5×10^{26} NO/flash. Furthermore, current evaluations of the global distribution of lightning (Goodman et al., 1988; Christian et al., 1992) in combination with estimated spatial distributions for different types of lightning (Orville, 1994) results in a nominal 100 global flashes/s being proportioned 75% to IC lightning and 25% to CG lightning. Seventy percent of CG lightning is associated with the tropics/subtropics, with 5% positive strokes, and 30% is associated with higher latitudes having 30% positive strokes. Combining these estimates with the midrange value for NO_x production per flash for each lightning type results in a conservative estimate for total NO_x production of 2.5 Tg N/yr by IC lightning, 3 Tg N/yr by negative CG lightning, and 1 Tg N/yr by positive CG lightning.

These global lightning estimates are found to be in generally good agreement with other independent NO_x assessments not dependent on the mechanistic details of lightning. For example, Albritton et al. (1984) constrained the global lightning source strength by examining nitrate deposition records from remote global areas that were expected to be free of impacts from anthropogenic sources. They estimated a lightning source of approximately 8 Tg N/yr (range 2 to 20). More recently, Levy et al. (1996) have used a global chemical transport model in conjunction with remote, upper tropospheric NO measurements to constrain all lightning sources. Their results, which critically depend on their choice of parameterizations for deep convection, indicate a range of values for NO_x production from lightning of 2 to 6 Tg N/yr.

A final issue of importance concerning lightning emissions relates to the altitude distribution of emissions. This is influenced not only by the initial production from lightning but also by subsequent vertical mixing in convective storms. Pickering et al. (1998) have recently produced estimates for the vertical distribution of lightning emissions. Their results show that most lightning NO_x is delivered to the upper

troposphere; however, the more vigorous mixing of midlatitude continental storms leads to more downward transport of lightning NO_x than for maritime storms or tropical continental storms. They also showed that the peak NO_x in continental storms occurs at higher altitudes than for maritime storms.

Aircraft Emissions

Subsonic aircraft emissions represent the most quantitatively known direct source of NO_x in the upper troposphere. Aircraft engines use an extremely efficient, high temperature combustion process that primarily produces CO_2, H_2O, and a few percent of other compounds. Like lightning, the initial NO_y content of these emissions consists primarily (>85%) of NO. Estimates of NO_x production are derived from assessments of the emission indices for various engines under different flight conditions and the annual amount of air traffic in terms of the kilograms of fuel consumed. The total strength of this source for both scheduled commercial and nonscheduled (e.g., military and chartered) air traffic in 1992 has been estimated at approximately 0.46 Tg N/yr for all altitudes with approximately 65% of the emissions occurring in the upper troposphere (>8 km), and, of that, approximately 45% has been assessed as occurring between 20° and 45°N with another 36% at latitudes further north (i.e., 55% of total emissions were at altitudes >8 km and latitudes > 20°N). This source of NO_x has also been projected to increase to approximately 1.3 Tg N/yr by the year 2015 (NASA, 1991). Because about 90% of the emissions occur in the free troposphere over the Northern Hemisphere, the impact from this source on the budgets and distributions of NO_x and ozone should be substantially different than in the Southern Hemisphere. In addition, this source may have a larger relative impact on upper tropospheric, winter time Northern Hemispheric NO_x distributions. This reflects the fact that during this period lightning and convection of surface NO_x emissions are both significantly reduced.

Stratosphere–Troposphere Exchange

The source of NO_y in the stratosphere is primarily N_2O oxidation by $O(^1D)$ in the upper stratosphere. Most of the mass that is transported into the troposphere, however, comes from the lower stratosphere. Peak levels of activity for this source occur primarily in the spring. This activity is found to be most vigorous near the subtropical and polar jet streams as well as in mid and high-latitude regions affected by atmospheric overturning associated with largc-scalc low-pressure disturbances and frontal systems. Even though the NO_x flux estimates are small, transported stratospheric odd nitrogen may still be a significant source of free tropospheric NO_x for some remote regions. For example, depending on how much of the NO_x production from lightning is transported to the stratosphere, estimates of the average global flux of NO_y from the stratosphere range from approximately 0.3 to 1 Tg N/yr (Ko et al., 1986; Murphy and Fahey, 1994). These values are close to balancing the stratospheric production of NO_x from N_2O (Kasibhatla et al., 1991).

Although the stratospheric NO_y source seems small compared to boundary layer NO_y sources, it is comparable in magnitude to other free tropospheric sources such as emissions from subsonic aircraft. Unlike subsonic aircraft emissions and lightning, which predominantly release NO directly into the free troposphere, input of NO_y from the middle stratosphere should primarily consist of HNO_3 with only a small, approximately 25% or less (i.e., 0.25 Tg N/yr), contribution from NO_x (Russell et al., 1988; Notholt et al., 1995). However, because removal of NO_y in the upper troposphere is extremely inefficient compared to that in the lower atmosphere, the NO_y from the stratosphere may have a long enough lifetime to impact the distribution of NO_x through recycling reactions, thereby influencing ozone production in the free troposphere.

Oceans

While this source is expected to be small, it could have greater importance for specific regions given its location far from the more dominant continental sources. Zafirou and McFarland (1981) conducted measurements in the equatorial Pacific and found that nitrite photolysis may provide a source of NO from the ocean. Based on air–sea exchange models and the difference between $[NO]_{ocean}$ and $[NO]_{air}$, several investigators derived a source strength of about 0.5 Tg N/yr (Logan, 1983; Liu et al., 1983; Torres and Thompson, 1993). This value is highly uncertain, however, since the data from the equatorial Pacific represent a small sample from a more biologically productive region of the ocean.

Ammonia Oxidation

A poorly quantified, but still potentially important source of NO_x is the oxidation of atmospheric ammonia (NH_3). A major uncertainty regarding this source is the lack of information on the tropospheric distribution of NH_3. NH_3 is initially oxidized by OH to form NH_2. NH_2 may go on to form NO_x through reaction with O_3, however, it may also react with NO_x to form N_2 or N_2O. Based on differences in rate coefficients, NH_3 oxidation should provide a net source of NO_x when ambient NO_x is less than 200 to 500 ppt, a condition that is prevalent in the remote troposphere. Recently, boundary layer NH_3 mixing ratios in the 50 to 900 ppt range (median 250 pptv) have been found over large stretches of the South Pacific and the Southern Ocean (J. Bradshaw, unpublished data). A reasonable estimate for this source is about 0.6 Tg N/yr, assuming a background tropospheric NH_3 value of 150 pptv and a 4-month lifetime for oxidation via OH.

4 TROPOSPHERIC DISTRIBUTION OF REACTIVE NITROGEN

Knowledge concerning the tropospheric distribution of NO_x is critical given its importance to ozone photochemistry. Over much of the remote atmosphere, NO concentrations hover near the critical level necessary for net photochemical produc-

tion of ozone. This critical NO level varies from as low as about 5 pptv to near 20 pptv depending on ambient conditions (Crawford et al., 1997). As a general rule, observations have shown NO to typically fall below critical levels over remote marine boundary layer environments away from NO_x sources where the lowest ozone values in the atmosphere also occur. By contrast, ozone production in the upper troposphere appears to be ubiquitous given observations of NO consistently above the critical level.

Despite the pivotal role NO_x plays in tropospheric photochemistry, current knowledge of its tropospheric distribution is based on very limited data, especially for remote regions. Reliable methods for measuring NO over the full range of its tropospheric variability (a few pptv to tens of ppbv) have been available since the late 1970s. Even so, field measurements of NO_x from ground, ship, and aircraft platforms provide only limited spatial and temporal coverage. Nevertheless, these observations do reveal some basic features in the global distribution of NO_x (Emmons et al., 1997; Bradshaw et al., 2000). For instance, gradients in NO_x observations are greatest near the surface. NO_x in urban areas is typically in the parts-per-billion range and a few hundred parts-per-trillion are common even in rural areas. For remote oceanic regions, however, NO_x levels are generally less than 50 ppt and often only a few parts per trillion. These trends in NO_x at the surface are consistent with most NO_x sources being land based and the short atmospheric lifetime for NO_x of a day or less. In the upper troposphere, gradients in NO_x are weaker owing to both the longer lifetime for NO_x and generally faster transport. NO_x values in the range of 50 to 200 pptv are often observed; however, observations ranging from only a few pptv to more than a ppbv of NO_x are not uncommon. These extremes are most likely due to the convection of NO_x-poor air in marine environments contrasted by the convection of NO_x-rich polluted air with additional inputs from lightning in continental regions. As a consequence of the difference in NO_x gradients at the surface and high altitude, NO_x is generally observed to increase with altitude over remote locations.

Some of these trends can be seen in data collected from NASA's DC-8 aircraft (see Fig. 3). NO_x in this figure has been estimated from daytime measurements of NO (solar zenith angle $< 70°$) by assuming photochemical equilibrium conditions for NO_2. These data have been taken from the following field campaigns: Pacific Exploratory Mission (PEM)–West A (1991) (Hoell et al., 1996), Transport and Atmospheric Chemistry Near the Equator–Atlantic (TRACE-A, 1992) (Fishman et al., 1996), PEM–Tropics A (1997) (Hoell et al., 1999), and Subsonics Assessment Ozone and Nitrogen Oxides Experiment (SONEX, 1998) (Singh et al., 1999). These campaigns have focused on taking measurements to characterize the remote oceanic troposphere with an emphasis on the upper troposphere. While flown in different years, each campaign was conducted during the fall season (September–November).

Figure 3 shows data for the boundary layer (0 to 1 km), the lower free troposphere (1 to 6 km), and the upper troposphere (6 to 12 km). In general, NO_x over the Atlantic is greater than over the Pacific at all altitudes. This is due to a closer proximity to NO_x sources, which are predominantly land based. The seasonal nature of NO_x sources is also important to the elevated levels of NO_x over the South Atlantic since measurements were taken during the biomass burning season

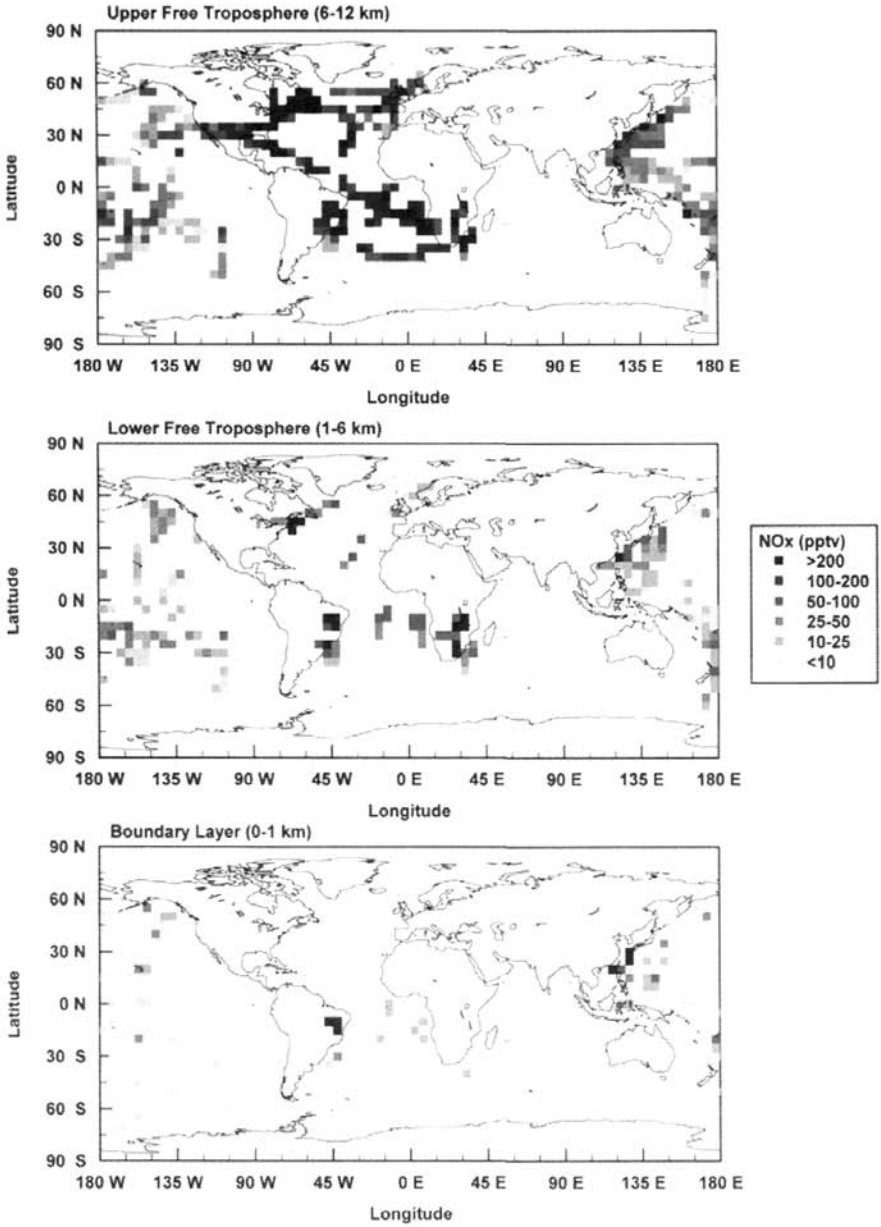


Figure 3 (see color insert) Distribution of NO_y based on measurements taken from NASA's DC-8 aircraft during fall (see text for details). Data are averaged on a $5^\circ \times 5^\circ$ latitude–longitude grid for three altitude ranges. See ftp site for color image.

for South America and Africa. While boundary layer data over and near continental areas are sparse, the strong gradient in surface NO_x is evident in the low values (typically < 10 pptv) over the South Pacific. Gradients are weaker at higher altitudes. The increase in NO_x with altitude over remote oceanic regions is also evident for both the South Atlantic and South Pacific data. The decrease in NO_x with altitude over continental areas is less evident since data is sparse, but NO_x values over South America, Southern Africa, and the South China coast do exhibit a decrease with altitude.

Information concerning the distribution of NO_y is even more limited than that for NO_x. The only NO_y species other than NO_x that have been measured with any regularity are HNO_3 and PAN. Total NO_y measurements have been more common, but there are still questions as to what these measurements represent since they often exceed expected values based on the sum of all NO_y constituents (Crosley, 1996). Measurements of HNO_3 in the remote troposphere have consistently fallen well below levels expected based on theory (Liu et al., 1992; Ridley et al., 1998; Schultz et al., 2000). This problem has been particularly troubling since theory predicts HNO_3 to be the dominant NO_y species in the remote troposphere. Heterogeneous mechanisms recycling HNO_3 to NO_x have been hypothesized as a potential solution (Fan et al., 1994; Chatfield, 1994; Hauglustaine et al., 1996; Lary et al., 1997). Possible underestimations in wet removal and partitioning of HNO_3 between gas and aerosol phases have been cited as well (Liu et al., 1992; Wang et al., 1998).

Measurements of PAN support the contention that it plays a strong role in sustaining NO_x in air masses as they are transported away from regions of strong industrial or biomass burning emissions. Decomposition of PAN was found to be adequate to explain NO_x observations at low altitude over eastern Canada (Fan et al., 1994), the South Atlantic (Jacob et al., 1996), and the western, North Pacific (Crawford et al., 1997). For even more remote regions, the decomposition of PAN in descending air masses can also be responsible for sustaining NO_x. This condition was observed by Schultz et al. (1999) over the remote South Pacific.

While global models can be used to estimate the distributions of NO_x and NO_y, the accuracy of these estimates is still very uncertain and are complicated by several factors. First is the level of uncertainty that remains for various NO_x sources, especially natural source strengths as well as spatial distributions. Second, there are still major questions concerning the recycling of NO_x from NO_y reservoir species and the potential role of aerosol in both the removal and recycling of NO_x. Finally, the wide range of photochemical lifetimes for NO_x requires atmospheric models to accurately represent small-scale transport processes (e.g., convective vertical transport of NO_x and wet deposition of HNO_3). The scales for these processes remain significantly smaller than the resolution of current photochemical transport models.

REFERENCES

Akimoto, H., and H. Narstu, Distributions of SO_2, NO_x, and CO_2 emissions from fuel combustion and industrial activities in Asia with $1^\circ \times 1^\circ$ resolution, *Atmos. Environ.*, *28*, 213–225, 1994.

Albritton, D. L., S. C. Liu, and D. Kley, Global nitrate deposition from lightning, in *Proceedings of the Conference on the Environmental Impact of Natural Emissions*, Air Pollution Control Association, Pittsburgh, PA, 1984, pp. 100–112.

Bradshaw, J., et al., Photofragmentation two-photon laser-induced fluorescence detection of NO_2 and NO: Comparison of measurements with model results based on airborne observations during PEM–Tropics A, *Geophys. Res. Lett.*, *26*, 471–474, 1999.

Bradshaw, J., D. Davis, G. Grodzinsky, S. Smyth, R. Newell, S. Sandholm, and S. Liu, Observed distributions of nitrogen oxides in the remote free troposphere from the NASA Global Tropospheric Experiment programs, *Rev. Geophys.*, *38*, 61–116, 2000.

Chatfield, R. B., Anomalous HNO_3/NO_x ratio of remote troposphere air: Conversion of nitric acid to formic acid and NO_x, *Geophys. Res. Lett.*, *21*, 2705–2708, 1994.

Christian, H. J., R. J. Blakeslee, and S. J. Goodman, Lightning imaging sensor (LIS) for the Earth Observing System, NASA Technical Memorandum 4350, Huntsville, AL, February 1992.

Corbett, J. J., P. S. Fischbeck, and S. N. Pandis, Global nitrogen and sulfur inventories for oceangoing ships, *J. Geophys. Res.*, *104*, 3457–3470, 1999.

Crawford, J., et al., An assessment of ozone photochemistry in the extratropical western North Pacific: Impact of continental outflow during the late winter/early spring, *J. Geophys. Res.*, *102*, 28469–28487, 1997.

Crosley, D. R., NO_y blue ribbon panel, *J. Geophys. Res.*, *101*, 2049–2052, 1996.

Emmons, L. K., et al., Climatologies of NO_x and NO_y: A comparison of data and models, *Atmos. Environ.*, *31*, 1851–1904, 1997.

Environmental Protection Agency (EPA), *Development of the 1980 NAPAP Emissions Inventory*, EPA/600/4-85-038, U.S. EPA, Research Triangle Park, NC, 1986, Chapter 4.

Fan, S. M., et al., Origin of tropospheric NO_x over subarctic eastern Canada in summer, *J. Geophys. Res.*, *99*, 16867–16877, 1994.

Fishman, J., J. M. Hoell, Jr., R. D. Bendura, R. J. McNeal, and V. W. J. H. Kirchoff, NASA GTE TRACE A experiment (September–October 1992): Overview, *J. Geophys. Res.*, *101*, 23865–23879, 1996.

Goldenbaum, G. C., and R. R. Dickerson, Nitric oxide production by lightning discharges, *J. Geophys. Res.*, *98*, 18333–18338, 1993.

Goodman, S. J., H. J. Christian, and W. D. Rust, A comparison of the optical pulse characteristics of intracloud and cloud-to-ground lightning as observed above clouds, *J. Appl. Meteorol.*, *27*, 1369–1381, 1988.

Hauglustaine, D. A., B. A. Ridley, S. Solomon, P. G. Hess, and S. Madronich, HNO_3/NO_x ratio in the remote troposphere during MLOPEX 2: Evidence for nitric acid reduction on carbonaceous aerosols? *Geophys. Res. Lett.*, *23*, 2609–2612, 1996.

Hoell, J. M., D. D. Davis, D. J. Jacob, M. O. Rodgers, R. E. Newell, H. E. Fuelberg, R. J. McNeal, J. L. Raper, and R. J. Bendura, Pacific Exploratory Mission in the tropical Pacific: PEM–Tropics A, August–September 1996, *J. Geophys. Res.*, *104*, 5567–5583, 1999.

Hoell, J. M., D. D. Davis, S. C. Liu, R. Newell, M. Shipham, H. Akimoto, R. J. McNeal, R. J. Bendura, and J. W. Drewry, Pacific Exploratory Mission—West A (PEM—West A): September–October 1991, *J. Geophys. Res.*, *101*, 1641–1653, 1996.

Jacob, D. J., and S. C. Wofsy, Budgets of reactive nitrogen, hydrocarbons, and ozone over the Amazon forest during the wet season, *J. Geophys. Res.*, *95*, 16737–16754, 1990.

Jacob, D. J., et al., Origin of ozone and NO_x in the tropical troposphere: Photochemical analysis of aircraft observations over the South Atlantic Basin, *J. Geophys. Res.*, *101*, 24235–24250, 1996.

Kasibhatla, P. S., H. Levy II, W. J. Moxim, and W. L. Chameides, The relative impact of stratospheric photochemical production on tropospheric NO_y levels: A model study, *J. Geophys. Res.*, *96*, 18631–18646, 1991.

Ko, M. K. W., M. B. McElroy, D. K. Weisenstein, and N. D. Sze, Lightning: A possible source of stratospheric odd nitrogen, *J. Geophys. Res.*, *91*, 5395–5405, 1986.

Kumar, P. P., G. K. Manohar, and S. S. Kandalgaonkar, Global distribution of nitric oxide produced by lightning and its seasonal variation, *J. Geophys. Res.*, *100*, 11203–11208, 1995.

Lary, D. J., A. M. Lee, R. Toumi, M. J. Newchurch, M. Pirre, and J. B. Renard, Carbon aerosols and atmospheric photochemistry, *J. Geophys. Res.*, *102*, 3671–3682, 1997.

Lawrence, M. G., W. L. Chameides, P. S. Kasibhatla, H. Levy II, and W. Moxim, Lightning and atmospheric chemistry: The rate of atmospheric NO production, in H. Volland (Ed.), *Handbook of Atmospheric Electrodynamics*, Vol. 1, CRC Press, Boca Raton, FL, 1995, pp. 189–202.

Lawrence, M. G., and P. J. Crutzen, Influence of NO_x emissions from ships on tropospheric photochemistry and climate, *Nature*, *402*, 167–170, 1999.

Lee, D. S., et al., Estimations of global NO_x emissions and their uncertainties, *Atmos. Environ.*, *31*, 1735–1749, 1997.

Levy, II, H., W. J. Moxim, and P. S. Kasibhatla, Global 3-dimensional time-dependent lightning source of tropospheric NO_x, *J. Geophys. Res.*, *101*, 22911–22922, 1996.

Liaw, Y. P., D. L. Sisterson, and N. L. Miller, Comparison of field, laboratory, and theoretical estimates of global nitrogen fixation by lightning, *J. Geophys. Res.*, *95*, 22489–22494, 1990.

Liu, S. C., et al., A study of the photochemical and ozone budget during the Mauna Loa Observatory Photochemistry Experiment, *J. Geophys. Res.*, *97*, 10463–10471, 1992.

Liu, S. C., M. McFarland, D. Kley, O. Zafirou, and B. Huebert, Tropospheric NO_x and O_3 budgets in the equatorial Pacific, *J. Geophys. Res.*, *88*, 1360–1368, 1983.

Lobert, J. M., et al., Experimental evaluation of biomass burning emissions: Nitrogen and carbon containing compounds, in J. S. Levine (Ed.), *Global Biomass Burning: Atmospheric Climate and Biospheric Implications*, MIT Press, Cambridge, MA, 1991.

Logan, J. A., Nitrogen oxides in the troposphere: Global and regional budgets, *J. Geophys. Res.*, *88*, 10785–10807, 1983.

Lübkert, B., and S. DeTilly, The OECD—Map emission inventory for SO_2, NO_x and VOC in western Europe, *Atmos. Environ.*, *23*, 3–15, 1989.

Lübkert, B., and K. H. Zierock, European emission inventories—A proposal of international worksharing, *Atmos. Environ.*, *23*, 37–48, 1989.

Murphy, D. M., and D. W. Fahey, An estimate of the flux of stratospheric reactive nitrogen and ozone into the troposphere, *J. Geophys. Res.*, *99*, 5325–5332, 1994.

NASA, High speed research program/atmospheric effects of stratospheric aircraft (HSRP/AESA), Annual Report, 1991.

Neff, J. C., M. Keller, E. A. Holland, A. W. Weitz, and E. Veldkamp, Fluxes of nitric oxide from soils following the clearing and burning of a secondary tropical rain forest, *J. Geophys. Res.*, *100*, 25913–25922, 1995.

Notholt, J., A. Meier, and S. Peil, Total column densities of tropospheric and stratospheric trace gases in the undisturbed arctic summer atmosphere, *J. Atmos. Chem.*, *20*, 311–332, 1995.

Orville, R. E., Cloud-to-ground lightning flash characteristics in the contiguous United States: 1989–1991, *J. Geophys. Res.*, *99*, 10833–10841, 1994.

Pickering, K. E., Y. Wang, W.-K. Tao, C. Price, and J.-F. Müller, Vertical distributions of lightning NO_x for use in regional and global chemical transport models, *J. Geophys. Res.*, *103*, 31203–31216, 1998.

Ridley, B., et al., Measurements of NO_x and PAN and estimates of O_3 production over the seasons during Mauna Loa Observatory Photochemistry Experiment 2, *J. Geophys. Res.*, *103*, 8323–8339, 1998.

Russell III, J. M., et al., Measurements of odd nitrogen compounds in the stratosphere by the ATMOS experiment on Spacelab 3, *J. Geophys. Res.*, *93*, 1718–1736, 1988.

Schultz, M. G., et al., On the origin of tropospheric ozone and NO_x over the tropical South Pacific, *J. Geophys. Res.*, *104*, 5829–5844, 1999.

Schultz, M. G., D. J. Jacob, J. D. Bradshaw, S. T. Sandholm, J. E. Dibb, R. W. Talbot, and H. B. Singh, Chemical NO_x budget in the upper troposphere over the tropical South Pacific, *J. Geophys. Res.*, *105*, 6669–6679, 2000.

Singh, H. B., A. M. Thompson, and H. Schlager, SONEX airborne mission and coordinated POLINAT-2 activity: Overview and accomplishments, *Geophys. Res. Lett.*, *26*, 3053–3056, 1999.

Torres, A. L., and A. M. Thompson, Nitric oxide in the equatorial Pacific boundary layer: SAGA 3 measurements, *J. Geophys. Res.*, *98*, 16949–16954, 1993.

Wagner, J., R. A. Walters, L. J. Maiocco, and D. R. Neal, *Development of the 1980 NAPAP Emissions Inventory*, U.S. Environmental Protection Agency, Washington, DC, 1986.

Wang, Y., J. A. Logan, and D. J. Jacob, Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry 2. Model evaluation and global ozone budget, *J. Geophys. Res.*, *103*, 10727–10755, 1998.

Williams, E. J., et al., An intercomparison of five ammonia measurement techniques, *J. Geophys. Res.*, *97*, 11591–11611, 1992.

Yienger, J. J., and H. Levy II, Empirical model of global soil-biogenic NO_x emissions, *J. Geophys. Res.*, *100*, 11447–11464, 1995.

Yokelson, R. J., D. W. T. Griffith, and D. E. Ward, Open-path Fourier transform infrared studies of large-scale laboratory biomass fires, *J. Geophys. Res.*, *101*, 21067–21080, 1996.

Zafirou, O. C., and M. McFarland, Nitric oxide from nitrite photolysis in the central equatorial Pacific, *J. Geophys. Res.*, *86*, 3173–3182, 1981.

CHAPTER 5

CARBON MONOXIDE IN THE ATMOSPHERE

PAUL NOVELLI

Carbon monoxide (CO) is present in trace quantities in the atmosphere. Although first detected in the late 1940s using solar spectroscopic methods,[1,2] few measurements of CO were made during the period between the early 1950s and the mid-1960s. However, as chromatographic and related detection techniques were developed, discrete measurements of CO were made in many locations around the world. These provided considerable insight on global tropospheric distributions; most notable among these was the observation that CO concentrations generally decreased from north to south.[3–5] The significance of CO in atmospheric chemistry was recognized in 1971 when Levy,[6] and McConnell et al.[7] proposed a photochemically driven, radical chain reaction linking the tropospheric cycles of methane (CH_4), CO, nitric oxide and nitrogen dioxide (NO_x), and formaldehyde (CH_2O), with those of the oxidants ozone (O_3), the hydroxyl (OH) and hydroperoxyl (HO_2) radicals. These models describe an atmosphere in which the photolysis of O_3 ($h\nu < 320$ nm) leads to the formation of OH, initiating a series of oxidation/reduction reactions that both produce and destroy CO, CH_2O, OH and HO_2.

In much of the background atmosphere the reaction of CO and OH [Eq. (1)] accounts for 90 to 95% of the loss of CO[8] and about 75% of the removal of OH.[9]

$$CO + OH + O_2 \rightarrow CO_2 + HO_2 \qquad (1)$$

While the stoichiometric relationship between CO oxidation and OH loss is dependent upon several possible reaction pathways, the inverse relationship between CO and OH concentrations suggested by Eq. (1) is expected in the background atmosphere.[9,10] Not only does the hydroxyl radical regulate the concentration of

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

CO, but oxidation at the expense of OH is also the primary removal pathway for many other reduced gases, several of which are radiatively important [e.g., CH_4, the hydrogenated chlorofluorocarbons (CFCs)]. Therefore, trends in atmospheric CO levels are expected to have an effect on climate through its role in regulating [OH], which in turn affects the levels of several important greenhouse gases.[11]

Carbon monoxide impacts both local and regional air quality through its influence on ozone. In areas of relatively high NO_x levels (>5 to 10 pmol/mol), such as urban areas or air parcels affected by fossil fuel or biomass burning, HO_2 produced through CO oxidation enters into a series of photochemical reactions that produce O_3:

$$NO + HO_2 \rightarrow NO_2 + OH \quad (2)$$

$$NO_2 + h\nu \rightarrow NO + O \quad (3)$$

$$O + O_2 \rightarrow O_3 \quad (4)$$

In the background atmosphere, where $[NO_x]$ is often < 5 pmol/mol, HO_2 produced by the oxidation of CO may destroy O_3.

$$O_3 + HO_2 \rightarrow 2O_2 + OH \quad (5)$$

As a result, the oxidizing capacity of the lower atmosphere is coupled to the concentrations, distributions, and trends of CO.

1 MEASUREMENT TECHNIQUES

Analytical Methods

Measurements of atmospheric CO are conducted using a variety of techniques. Solar spectra recorded at 4.7 μm are used to derive total column abundances and column-averaged mixing ratios.[12,13] Nondispersive infrared radiometry (NDIR)[14,15] and tunable diode laser spectroscopy (TDLS)[16] also make use of CO absorption at 4.7 μm. Both techniques provide a continuous measurement of CO; however, the TDLS provides greater precision (1 ppb) and a higher measurement frequency (10 s), about a factor of 10 greater than NDIR. Gas chromatography (GC), when coupled with a number of different detectors can provide high precision and low detection limits.[17–19] The most common detectors used with GC are flame ionization (with prior conversion of $CO + H_2 \rightarrow CH_4$), electron capture, and hot mercuric oxide reduction. GC techniques can provide a high precision (1 ppb) with a discontinuous measure of CO (frequency on the order of a few minutes).[20]

Calibration

The gas chromatographic methods, NDIR, and some TDLS techniques require calibration against samples with known gas amounts. As far as we are aware,

there is no national or commercial laboratory that provides certified CO reference gases at levels found in the background atmosphere. Groups measuring CO must therefore dilute high concentration certified gases to atmospheric levels or obtain standards from other laboratories. Laboratory intercomparisons of the reference gases used by various researchers have shown large differences between groups (up to 25 to 50%).[21–23] CO standards may also be subject to drift over time. Efforts have been made since the early 1990s to compare CO measurements in both the laboratory and the field.[14,21] However, differences between groups still exist, and the integration of data sets requires some prior understanding of how the measurements compare.

2 GLOBAL CO DISTRIBUTIONS

Surface CO

Background Atmosphere. CO varies both temporally and spatially. Figure 1 presents a smoothed representation of the surface distribution of CO in the background marine boundary layer (MBL) as a function of latitude and time. The surface illustrates that CO mixing ratios in both hemispheres exhibit seasonal variation, and

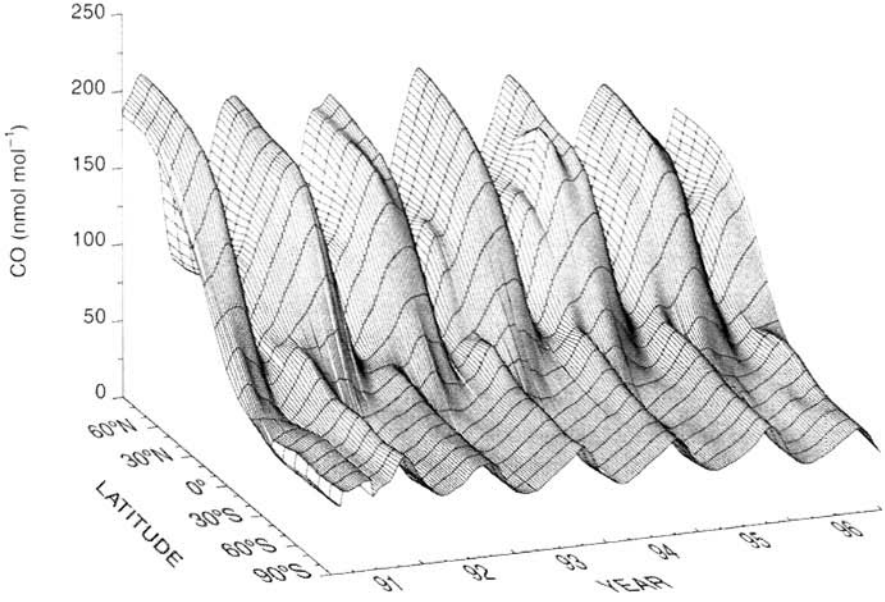


Figure 1 Smooth surface representing the distribution of CO in the marine boundary layer. The surface was created from 38 time series determined from sampling locations in the NOAA/CMDL Cooperative Air Sampling Program. CO mixing ratios were combined in 5° latitude bands, longitudinal differences were averaged, and the combined time series were smoothed in both time and space.[24]

although there are considerable interannual variations, repeatable patterns occur from year to year. Most notable is the seasonal cycle and the interhemispheric gradient. Greatest CO mole fractions in the MBL [200 to 225 nmol CO/mol air (ppb)] are found in the high latitudes of the Northern Hemisphere during late winter/early spring. The high Northern Hemisphere also exhibits the greatest seasonal amplitude (120 to 140 ppb). The imbalance of sources in winter (mostly anthropogenic pollution from the midlatitudes) and sink (when OH levels are lowest) leads to an accumulation of CO in the high Northern Latitudes. Lowest CO mixing ratios in the boundary layer (40 to 50 ppb) are found during the southern summer, where low concentrations are further depressed by reaction with OH. The interhemispheric gradient also exhibits a strong seasonality. The largest difference between the high northern and high southern hemispheres (~150 ppb) occurs in February/March and the minimum difference (10 to 20 ppb) occurs in September/October.[17,18,24]

Polluted Atmosphere. CO levels in urban locations and areas of regional-scale pollution are greater than those found in the background atmosphere, with CO mixing ratios in urban areas often reaching ppm level, orders of magnitude greater than those found in the background troposphere. CO is defined as a criteria species for urban pollution. The lifetime of CO is on the order of several months, and emissions can be transported far from the original source region.[25] Even in areas far distant from CO sources, wide-scale, diffuse pollution may enhance CO levels (up to twice background levels). Air parcels downwind of areas where combustion occurs can also show elevated levels of O_3.[26,27] The enhanced O_3 often reflects its photochemical production [Eqs. (2)–(4)], which is favored in environments having both high CO and NO_x.[28]

Free Troposphere

Near the planet's surface, CO varies with season, proximity to source regions, and latitude. Above the boundary layer, and in the middle and upper troposphere, CO also shows seasonal cycles and spatial distributions. At altitudes higher than a few kilometers, mixing ratios are largely determined by surface source distributions and by vertical and horizontal transport.[29,30] Mixing ratios determined in the free troposphere at mountain observatories in the Northern Hemisphere are typically lower than measurements made at sea level at similar times and latitudes.[18] CO mixing ratios in the free troposphere, studied from aircraft, show interhemispheric differences similar to those at the surface (higher levels in the north compared to the south).[30,31] In the Northern Hemisphere, CO often decreases with height,[31] reflecting the abundance of surface sources, but this may be seasonally dependent.[32] In the Southern Hemisphere, CO may increase with altitude or remain relatively constant,[31] and strong convective transport in the tropics can deliver CO to the middle and upper troposphere. Across the tropopause, CO mixing ratios fall below 50 ppb.[33] Transport from the stratosphere brings air with low CO into the troposphere.

Satellite Measurements

Global distributions of CO in the middle troposphere have been determined by the Measurement of Air Pollution from Satellite (MAPS) instrument. MAPS uses a nadir viewing gas filter correlation radiometry with a maximum signal between 400 and 300 mbar and has been flown aboard the U.S. space shuttle four times between 1981 and 1994. Results obtained in October 1984 and 1994 show very high levels of CO over the southern tropics,[25,34] evidence of the strong effect the transport of emissions from surface biomass burning can have on the middle troposphere.

Future measurements from space promise to provide long-term global coverage of CO distributions in the troposphere. The Measurement of Pollution in the Troposphere (MOPITT) instrument was launched December 1999 aboard the EOS *TERRA* (previously known as the *AM-1*) satellite. MOPITT, like MAPS, is a gas filter radiometer that will determine the column abundance of CO. In addition, MOPITT also will retrieve tropospheric profiles of CO (at 4.7 μm) through pressure and length modulation of the correlation cell. Total column abundances of CO and CH_4 will also be measured (at 2.3 μm).[35] MOPITT is expected to provide nearly continuous monitoring of tropospheric CO for a period of at least 5 years. The EOS *Aura* satellite (formerly denoted as *CHEM-1*) is scheduled for launch in June 2003. The *Aura* payload will include TES (tropospheric emission spectrometer), an infrared imaging Fourier transform spectrometer with high spectral resolution that will determine global distributions of CO (and other radiatively trace gases) in the troposphere and lower stratosphere (*http://aura.nasa.gov/tes*).

3 GLOBAL CO BUDGET

Tropospheric distributions of CO reflect its sources and sink combined with the effects of transport. There are believed to be four major sources of CO (Table 1): fossil fuel combustion and industrial activities, biomass burning, the oxidation of methane, and the oxidation of nonmethane hydrocarbons, primarily isoprene and the monoterpenes.[8,36] Anthropogenic activities are thought to account for about two-

TABLE 1 Estimated Sources and Sinks of CO Typical of Last Decade[36]

Sources	Range (Tg CO/yr)	Sinks	Range (Tg CO/yr)
Industry and transportation	300–500	OH reaction	1400–2600
Biomass burning	300–700	Soil uptake	250–640
Emissions from vegetation	60–160	Loss to the stratosphere	~100
Oceans	20–200		
CH_4 oxidation	400–1000		
NMHC oxidation	200–600		
Total sources	1280–3160	Total sinks	1750–3340

thirds of the total source, and reaction with OH radicals is responsible for much of the loss of CO. Sources are unevenly divided between the hemispheres, with as much as 95% of the fossil fuel source, 63% of the biomass burning source, and 68% of the production from the oxidation of NMHC occurring in the Northern Hemisphere.[8] Three-dimensional global transport models, using sources of these magnitudes, have reproduced the measured surface distributions and seasonal cycles with varying degrees of success.[37–39]

4 TROPOSPHERIC TRENDS

An excess of sources relative to sinks leads to accumulation of gases in the atmosphere. Increasing sources through the industrial era have enhanced atmospheric burdens of CO_2 and CH_4.[36] Similar long-term increases in CO could be expected; however, the few analyses of CO in firn and ice samples have not produced convincing evidence of such a change.[39]

Time series of CO mixing ratios often show periods of increase and decrease.[17,24] Spectroscopic measurements made at Jungfrauhoch, Switzerland, during 1950–1951 and again in 1985–1987 suggested an average rate of increase of ~1% per year in the total column abundance of CO above the European boundary layer.[12] A similar rate of increase was seen in column measurements made over western Russia.[40] Surface measurements made at six sites (evenly distributed between the Northern and Southern Hemispheres) during 1981–1986 suggested a similar rate of increase.[41] In contrast, no significant trend could be identified at Cape Point, South Africa, during the period from the early 1970s through the mid-1980s.[18,42] Trends largely reflect imbalances in its sources and sinks. The reported long-term CO increase in the Northern Hemisphere has been attributed to increasing CO emissions from industrial and transportation-related sources.[12,39,43] However, a quantitative study relating CO emissions and increased atmospheric mixing ratios is still needed.

The long-term increase in CO may have slowed, then reversed in the late 1980s. Khalil and Rasmussen[41,43] present time series determined at six sites beginning 1981 that show an increase in CO over the period 1981–1986, followed by a decrease during 1987–1992 (Fig. 2). The absolute decline in the Northern Hemisphere was about twice that in the south; in the Southern Hemisphere the relative rate of decrease was four times that in the Northern Hemisphere. Novelli et al.[44] reported results from the NOAA/CMDL air sampling network showing a 10% decrease in global average CO mixing ratios during 1992–1993. And while CO declined in both hemispheres, the absolute rate of decrease in the Northern Hemisphere was nearly twice that in the south, while the relative rates were the same (approximately 6 to 7% per year). After 1993, CO levels showed short periods of increase and decrease with some recovery toward pre-1992 levels.[24]

CO time series determined in the north show a significant decrease over the past 10 years; in contrast, a trend in the Southern Hemisphere is more difficult to discern, due in part to the high level of interannual variability. This high variability is likely

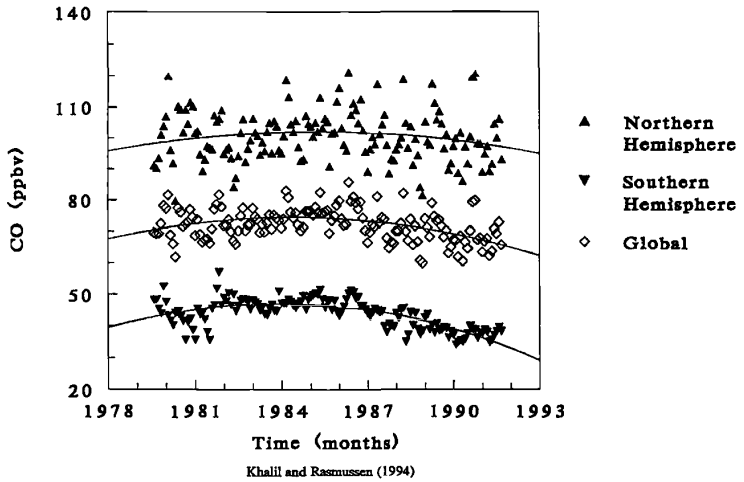


Figure 2 Time series of deseasonalized hemispheric and global mean CO mixing ratios. (Reprinted with permission from Khalil and Rasmussen.[43])

related to yearly variations in emissions from biomass burning. The short-term increases and decreases seen in the recent CO time series[17,24] may be related to interannual variability in biomass burning[43,44] and a short-term increase in OH related to the eruption of Mt. Pinatubo in June 1991.[39,46] Decreased emissions from anthropogenic sources in the Northern Hemisphere have contributed to the observed decrease[44,45].

REFERENCES

1. Migeotte, M., The fundamental band of carbon monoxide at 4. 7 u in the solar spectrum, *Phys. Rev.*, *75*, 1108–1109, 1949.
2. Adel, A., Identification of carbon monoxide in the atmosphere above Flagstaff, Arizona, *J. Astrophys.*, *116*, 442–443, 1952.
3. Robbinson, E. and R. C. Robbins, Atmospheric background concentrations of carbon monoxide, *Ann. New York Acad. Sci.*, *174*, 89–95, 1970.
4. Seiler, W., and C. Junge, Carbon monoxide in the atmosphere, *J. Geophys. Res.*, *75*, 2217–2226, 1970.
5. Seiler, W., and U. Schmidt, New aspects on CO and H_2 cycles in the atmosphere, in N. J. Derco and E. J. Trublar (Eds.), *Proceedings of the International Conference on the Structure, Composition, and General Circulation of the Upper and Lower Atmos. and Possible Anthropogenic Perturbations*, Association of Meteorological and Atmospheric Physics, Toronto, 1974.
6. Levy II, H., Natural atmosphere: Large radical and formaldehyde concentrations predicted, *Science*, *173*, 141–143, 1971.

7. McConnell, J. C., M. B. McElroy, and S. C. Wofsy, Natural sources of atmospheric CO, *Nature*, *233*, 187–188, 1971.
8. Logan, J. A., M. J. Prather, S. C. Wofsy, and M. B. McElroy, Tropospheric chemistry: A global perspective, *J. Geophys. Res.*, *86*, 7210–7254, 1981.
9. Thompson, A. M., The oxidizing capacity of the atmosphere: Probable past and future changes, *Science*, *256*, 1157–1165, 1992.
10. Sze, N. D., Anthropogenic CO emissions: Implications for the atmospheric CO-OH-CH_4 Cycle, *Science*, *195*, 673–674, 1977.
11. Daniel, J. S and S. Soloman, On the climate forcing of carbon monoxide, *J. Geophys. Res.*, *103*, 13249–13260, 1998 .
12. Zander, R., Ph. Demoulin, D. H. Ehhalt, U. Schmidt, and C. P. Rinsland, Secular increase of the total column abundance of carbon monoxide above central Europe since 1950, *J. Geophys. Res.*, *94*, 11021–11028, 1990.
13. Wallace, L., and W. Livingston, Spectroscopic observations of atmospheric trace gases over Kitt Peak, 2. Nitrous oxide and carbon monoxide from 1979 to 1985, *J. Geophys. Res.*, *95*, 16383–16390, 1990.
14. Doddridge, B. G., R. R. Dickerson, T. G. Spain, S. J. Oltmans, and P. C. Novelli, Measurements of carbon monoxide at Mace Head, Ireland, in ozone in the troposphere and the stratosphere, in R. D. Hudson (Ed.), *Proc. Quad. Ozone Symp.*, *1992*, NASA Conference Publication No. 3266, NASA, Greenbelt, MD, 1994, pp. 134–137.
15. Parrish, D. D., J. S. Holloway, and F. C. Fehsenfeld, Routine, continuous measurement of carbon monoxide with parts per billion precision, *Environ. Sci. Technol.*, *28*, 1615–1618, 1994.
16. Sachse, G. W., G. F. Hill, L. O. Wade and M. G. Perry, Fast-response, high-precision carbon monoxide sensor using a tunable diode laser absorption technique, *J. Geophys. Res.*, *92*, 2071–2081, 1987.
17. Brunke, E.-G., H. E. Scheel, and W. Seiler, Trends of tropospheric CO, N_2O and CH_4 as observed at Cape Point, South Africa, *Atmos. Environ.*, *24A*, 585–595, 1990.
18. Novelli, P. C., L. P. Steele, and P. P. Tans, Mixing ratios of carbon monoxide in the troposphere, *J. Geophys. Res.*, *97*, 20731–20750, 1992.
19. Hurst, D. F., P. S. Bakwin, R. C. Myers, and J. W. Elkins, Behavior of trace gas mixing ratios on a very tall tower in North Carolina, *J. Geophys. Res.*, *102*, 8825–8835, 1997.
20. Novelli, P. C., CO in the atmosphere: Measurement techniques and related issues, *Chemosphere*, *1*, 115–126, 1999.
21. Novelli, P. C., J. W. Elkins, and L. P. Steele, The development and evaluation of a gravimetric reference scale for measurement of atmospheric carbon monoxide, *J. Geophys. Res.*, *96*, 13109–13121, 1991.
22. Weeks, I. A., I. E. Galbally, P. J. Fraser, and G. Matthews, Comparison of the carbon monoxide standards used at Cape Grim and Aspendale, in B. W. Forgan and G. P. Ayers (Eds.), *Baseline Atmospheric Program, 1987*, Australian Government Department of Science and Technology, Canberra, Australia, 1989, pp. 21–25.
23. Novelli, P. C., V. S. Connors, H. G. Reichle, Jr., B. E. Anderson, C. A. M. Brenninkmeijer, E.-G. Brunke, B. G. Doddridge, V. W. J. H. Kirchhoff, J. K. S. Lam, K. A. Masarie, T. Matsou, D. D. Parrish, H. E. Scheel, and L. P. Steele, An internally consistent set of globally distributed atmospheric carbon monoxide mixing ratios developed using results from an intercomparison of measurements, *J. Geophys. Res.*, *103*, 19285–19293, 1998.

24. Novelli, P. C., K. A. Masarie, and P. M. Lang, Distributions and recent trends of carbon monoxide in the troposphere, *J. Geophys. Res.*, *103*, 19015–19033, 1998.

25. Reichle, H. G., Jr., V. S. Connors, J. A. Holland, R. T. Sherrill, H. A. Wallio, J. C. Casas, B. B. Gormsen, and W. Seiler, The distribution of middle tropospheric carbon monoxide during early October 1984, *J. Geophys. Res.*, *95*, 9845–9856, 1990.

26. Fishman, J., K. Fakharuzzaman, B. Cros, and D. Nganga, Identification of widespread pollution in the Southern Hemisphere deduced from satellite analyses, *Science*, *252*, 1693–1696, 1991.

27. Jaffe, D. et al., Transport of Asian air to North America, *Geophys. Res. Lett.*, *26*, 711–714, 1999.

28. Fishman, J. and P. J. Crutzen, The origin of ozone in the troposphere, *Nature*, *272*, 855–858, 1978.

29. Seiler, W., and J. Fishman, The distribution of carbon monoxide and ozone in the free troposphere, *J. Geophys. Res.*, *86*, 7255–7265, 1981.

30. Heidt, L. E., J. P. Krasnec, R. A. Lueb, W. H. Pollock, B. E. Henry, and P. J. Crutzen, Latitudinal distributions of CO and CH_4 over the Pacific, *J. Geophys. Res.*, *85*, 7329–7336, 1980.

31. Marenco, A., M. Macaigne, and S. Prieur, Meridional and vertical CO and CH_4 distributions in the background troposphere (70N–60S; 0–12 k altitude) from the scientific aircraft measurements during the Stratoz III experiment (June 1984), *Atmos. Environ.*, *23*, 185–200, 1989.

32. Yurganov, L. N., D. A. Jaffee, E. Pullman, and P. C. Novelli, Total column and surface densities of atmospheric carbon monoxide in Alaska, 1995, *J. Geophy. Res.*, *103*, 19337–19347, 1998.

33. Seiler, W., and P. Warneck, Decrease of the carbon monoxide mixing ratio at the tropopause, *J. Geophys. Res.*, *77*, 3204–3214, 1972.

34. Connors, V. S., B. B. Gormsen, S. Nolf, and H. G. Reichle, Jr., Spaceborne observations of the global distribution of carbon monoxide in the middle troposphere during April and October 1994, *J. Geophys. Res.*, *104*, 21455–21470, 1999.

35. Drummond, J. R., Measurements of pollution in the troposphere (MOPITT), in J. C. Gille and G. Visconti, (Eds.), *The Use of EOS for Studies of Atmospheric Physics*, North Holland, Amsterdam, 1992, pp. 77–101.

36. Intergovernmental Panel on Climate Change (IPCC), Climate Change 1994: Radiative Forcing of Climate Change, in J. T. Houghton, L. G. M. Filho, J. Bruce, H. Lee, B. A. Callander, E. Haites, N. Harris, and K. Maskell (Eds.), IPCC, University Press, Cambridge, England, 1995.

37. Allen, D. J., P. Kasibhatla, A. M. Thompson, R. B. Rood, B. G. Doddridge, K. E. Pickering, R. D. Hudson, and S.-J. Lin, Transport-induced interannual variability of carbon monoxide determined using a chemistry and transport model, *J. Geophys. Res.*, *101*, 28655–28669, 1996.

38. Granier, C., J.-F. Muller, S. Madronich, and G. P. Brasseur, Possible causes for the 1990–1993 decrease in the global tropospheric CO abundances: A three-dimensional sensitivity study, *Atmos. Environ.*, *30*, 1673–1682, 1996.

39. Haan, D., and D. Raynaud, Ice core record of CO variations during the last two millennia: Atmospheric implications and chemical interactions within the Greenland ice, *Tellus, 50B*, 253–262, 1998.

40. Yurganov, L. N., E. I. Grechko, and A. V. Dzhola, Zvenigorod carbon monoxide total column time series: 27 years of measurements, *Chemosphere*, *1*, 127–136, 1999.
41. Khalil, M. A. K., and R. A. Rasmussen, Carbon monoxide in the Earth's atmosphere: Indications of a global increase, *Nature*, *332*, 242–245, 1988.
42. Seiler, W., H. Geihl, E.-G. Brunke, and E. Halliday, The seasonality of the CO abundance in the Southern Hemisphere, *Tellus*, *36B*, 219–231, 1984.
43. Khalil, M. A. K., and R. A. Rasmussen, Global decrease in atmospheric carbon monoxide concentration, *Nature*, *370*, 639–641, 1994.
44. Novelli, P. C., K. A. Masarie, P. P. Tans, and P. M. Lang, Recent changes in atmospheric carbon monoxide, *Science*, *263*, 1587–1590, 1994.
45. Bakwin, P. S., P. P. Tans, and P. C. Novelli, Carbon monoxide budget in the Northern Hemisphere, *Geophys. Res. Lett.*, *21*, 433–436, 1994.
46. Bekki, K. S., K. S. Law, and J. A. Pyle, Effect of ozone depletion on atmospheric CH_4 and CO concentrations, *Nature*, *371*, 595–597, 1994.

CHAPTER 6

ATMOSPHERIC METHANE

M. A. K. KHALIL AND M. J. SHEARER

1 INTRODUCTION

Methane has been increasing in the atmosphere for about two centuries, resulting in current concentrations that are more than twice the natural levels. Because of these trends, methane is considered to be a potentially important contributor to global warming and other man-made environmental changes that may occur in the future. As a greenhouse gas, every gram of methane released to the atmosphere is about 20 to 60 times as effective as a gram of CO_2, when considered over periods of 20 to 50 years. Moreover, methane has other critical roles in atmospheric chemistry that are also affected by its trends. It exercises a strong influence on the abundance of hydroxyl radicals (OH). These radicals in turn are responsible for removing many man-made and natural gases from the atmosphere. Increasing levels of methane can lead to a lowering of OH levels, that could in turn lead to increases of other gases that may be undesirable. Methane has a complex role in stratospheric chemistry where it is a source of water vapor that tends to deplete the ozone layer but, on the other hand, methane can scavenge chlorine atoms thus protecting the ozone layer from destructive effects of man-made chlorofluorocarbons. In this role high levels of methane are considered desirable. Methane is, therefore, integrally involved in the stability of Earth's environment and, as such, is regarded as one of the important trace gases that are significantly affected by human activities.

We will start by examining the observational data consisting of global distributions and trends. The atmospheric observations are the foundation for most of our current knowledge of the global cycle of methane and our interest in its possible environmental effects. Next we will see how these observations are explained in terms of the processes that produce and destroy methane. Based on the understand-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

a.
b.
Concentration CH4 (ppbv)
Year
1800
1600
1400
1200
1000
800
600
1000
1200
1400
1600
1800
2000
1795
1835
1875
1915
1955
1995

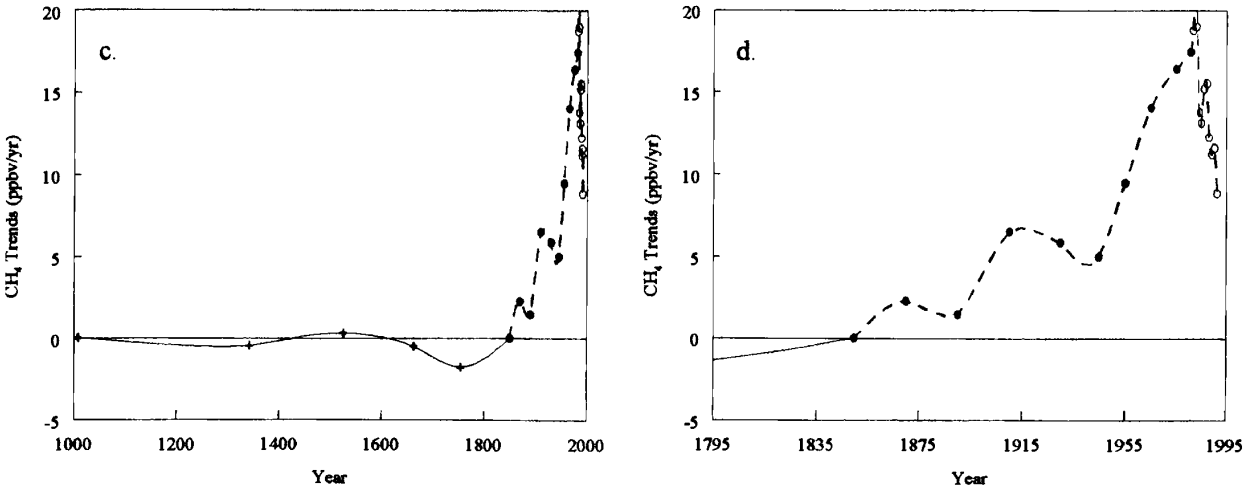


Figure 1 Concentrations of methane (*a*) over the last 1000 years and (*b*) over the last 200 years. Data of Etheridge et al. (1992) (•) and Rasmussen and Khalil (1984) (+) are from ice core samples. Data of Khalil and Rasmussen (○) are global averages from weekly flask samples collected at various latitudes. Trends of methane concentrations (*c*) during the last 1000 years and (*d*) over the last 200 years are calculated from the data shown in (*a*) and (*b*). Rapid increases in methane started only about 200 years ago. These are linear regression estimates of trends over various (nonoverlapping) periods of time between about A.D. 0 and the present. For data between 1840 and 1940, trends were calculated over 20-year periods; for 1940–1980, over 10-year periods; and for 1980–1992 over 2-year periods. The calculated trends were placed at the middle of the time span in each calculation. The earlier data are more sparse. Trends for the period between A.D. 0 and 1800 were calculated for every 10 data points, and the trends are placed at the average time spanned by the 10 data points.

ing gained by such a discussion, we can evaluate plausible expectations for future concentrations and the resulting environmental impact of methane.

2 ATMOSPHERIC OBSERVATIONS

Atmospheric concentrations of methane have been measured systematically for nearly 20 years (Rasmussen and Khalil, 1981; Khalil and Rasmussen, 1983, 1990a; Blake and Rowland, 1988; Steele et al., 1992; Khalil et al., 1993; Dlugokencky et al., 1994). There is an additional record from many independent measurements spanning another 15 years or so back to the early 1960s (Khalil et al., 1989). For earlier times, the ice core record is the only source of information. It extends back over 150,000 years, but for our interest here only the last 1000 years or so are important (Rasmussen and Khalil, 1984; Chappellaz et al., 1990; Etheridge et al., 1992). This record is summarized in Figures 1 and 2. The first panel of these figures shows the time history of methane over the last 1000 years, 100 years, and the most recent decades, and the second panel shows the trends of methane over the same periods.

These data establish two important results: First, that methane concentrations have increased by a factor of about 2.5 over the last 100 to 200 years. And second, that the rates of increase reached peak values during the 1980s, but have been declining since. The rapid increases observed in the 1980s suggested that methane could contribute significantly to global warming in the future. These observations were the compelling reason that drove much of the research on a systematic study of the cycle of atmospheric methane. Later we will return to why these trends are changing.

The most recent decades of data shown in Figure 1 contain many features that reflect the production and destruction processes of methane. There are two salient patterns: The seasonal cycles and the latitudinal concentration gradient. These features are shown more clearly in Figures 3 and 4, respectively.

The data shown are taken at Earth's surface at locations that are far from local sources. As such these concentrations and their patterns represent the large-scale distribution of methane in the atmosphere. In the vertical, up to the tropopause, methane mixing ratios remain nearly the same, implying that the actual concentration of methane in molecules/cm^3 falls off at the same rate as the density of air or approximately 12.5%/km. At higher altitudes, in the stratosphere, the concentrations (molecules/cm^3) fall at an approximate rate of 5%/km as shown in Figure 5.

These observations provide qualitative evidence for several important conclusions regarding the global cycle of methane. Clearly, the methane cycle is out of balance when considered over decadal time scales as evidenced by the generally increasing trends. Moreover, this imbalance arose over the last 100 to 200 years since, before that time, the concentration was unchanging, at least over the previous 1000 years. This would mean that in recent times, more methane is put into the atmosphere than is being removed annually. Second, the latitudinal gradient suggests that the production of methane is considerably higher in the Northern Hemisphere compared with the Southern Hemisphere, if we assume that the destruction processes are similar in

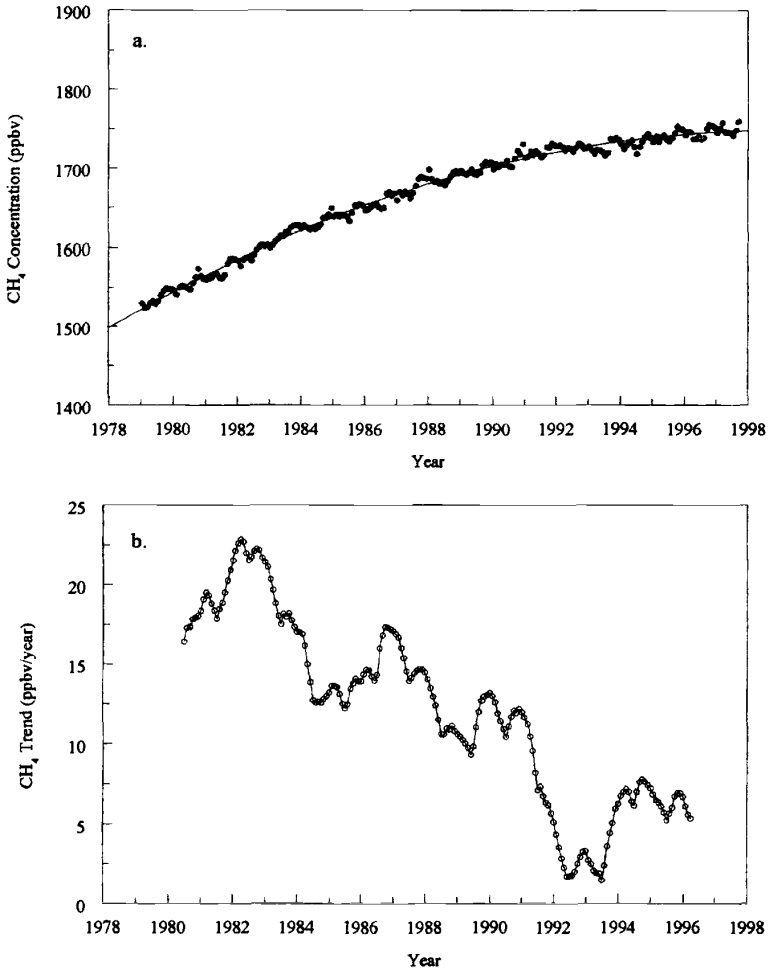


Figure 2 Global average concentration of methane from weekly flask samples collected from six different sites (*a*). The trend of methane (*b*) is calculated by linear regression of 3-year overlapping periods of time, plotted at the center point of the time period.

the two hemispheres. Both these observations, and the timing of the increasing trends, suggest that these changes are caused by human activities. Later, we will look at more direct evidence for this conclusion. Finally, the seasonal cycle suggests that at middle and higher latitudes, the imbalance between production and destruction is greater during winters than summers. This, we will find later, is consistent

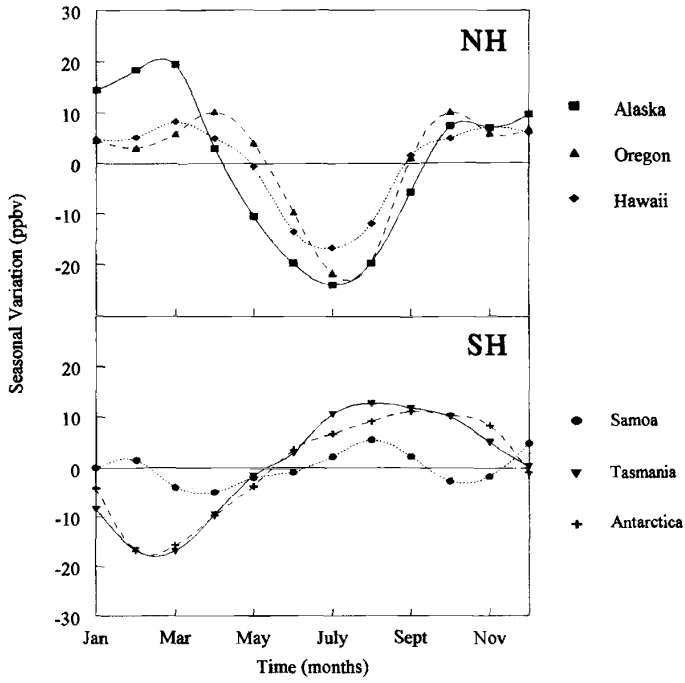


Figure 3 Average seasonal variations of CH_4 at six sites.

with the idea that methane is destroyed by reactions in the atmosphere with hydroxyl radicals that are produced by photochemical processes and hence are most abundant during summers at middle and higher latitudes. While complete objectivity would allow a few alternate explanations, these serve as good working hypotheses. An understanding of the production and destruction processes of methane should explain these observations quantitatively. We will aim toward this goal in the remainder of this chapter.

3 MASS BALANCE

The mass balance of a gas in a hypothetical infinitesimal volume of the atmosphere, in a unit of time, can be expressed as the production less the destruction, added to the net transport of the gas. The net transport can either increase or reduce the concentration within this box during the time of interest. If these three components are perfectly balanced, then the concentration will remain constant; if not, the concentration will change. A fuller treatment of the mass balance requires taking

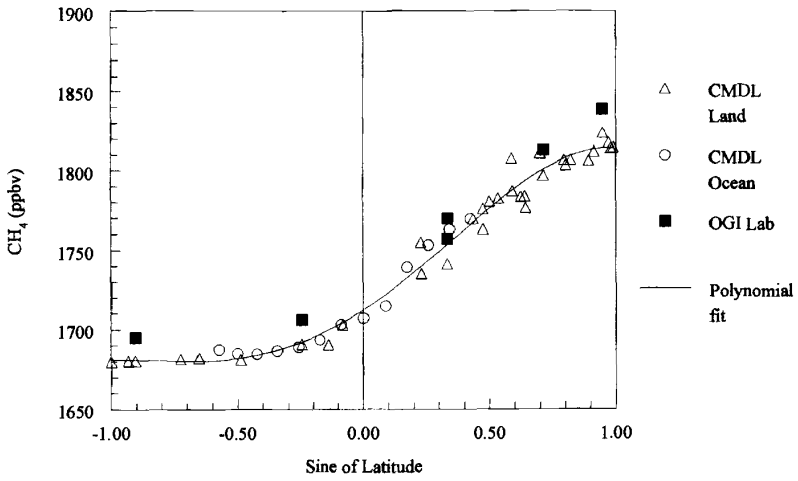


Figure 4 Latitudinal distribution of CH_4. The 1996 NOAA/CMDL data (Dlugokencky et al., 1994) are shown next to the Rasmussen and Khalil flask sampling data from six sites. (Khalil and Rasmussen, 1983) (Calibration difference: $C_{NOAA} = C_{R\&K} - 12$ ppbv.)

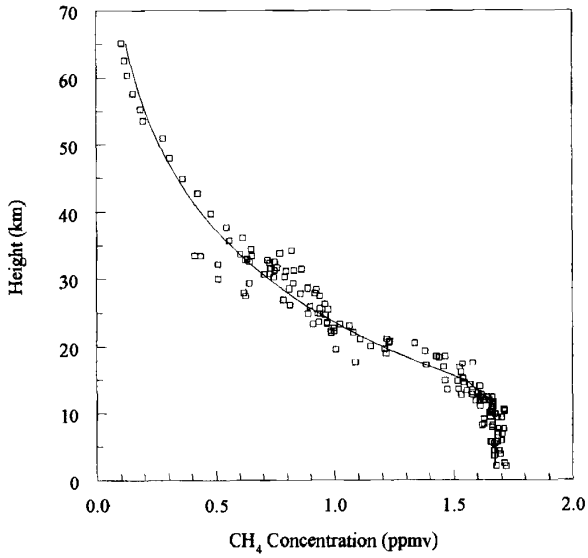


Figure 5 Vertical distribution of CH_4 at 44°N latitude. Data from Fabien et al. (1981), Schmidt et al. (1984, 1987), and Taylor et al. (1989). Concentrations adjusted to base year 1990.

these factors into consideration for each point in the atmosphere and for each unit of time. For our purposes here we will deal with a simplified concept whereby the entire atmosphere is regarded as a single reservoir into which we put methane from its sources and from which methane is removed by a series of chemical and physical processes. Moreover, we take each loss process to be proportional to the amount of methane present at any given time. Considerations of transport of methane are no longer explicitly needed since all methane stays within the atmosphere, and what is moved from one part by the winds, goes to another location, still within the global box, thus not affecting the amount of methane in the global atmosphere. This model can be stated as:

$$\frac{dC}{dt} = S - \frac{C}{\tau} \tag{1}$$

Here C is the amount of methane in Tg (1 Tg $= 10^{12}$ g), S is the emissions from all sources in Tg/yr, and τ is the effective atmospheric lifetime in years. τ is a composite lifetime due to all the processes that remove methane from the atmosphere, or $1/\tau = 1/\tau_1 + 1/\tau_2 + \cdots + 1/\tau_N$ where $\tau_1, \tau_2, \ldots, \tau_N$ represent the lifetimes due to each of N processes. For direct comparisons with measurements C can be converted to ppbv and hence S is expressed in ppbv/yr. The conversion factor is 1 Tg $\approx$ 2.8 ppbv in the global atmosphere.

In the mass-balance equation, we know the solution (C) based on the atmospheric measurements, so our task is to find the two remaining parts of the budget, namely the emissions (S) or the lifetime (τ). For the case of methane, the lifetime is calculated independently, so that the mass-balance equation is essentially a tool for finding the sources that have combined emissions that are consistent with the measured concentrations and calculated loss rates [Eq. (1)]. How Eq. (1) is used will be discussed next; then we will show the recent budgets consistent with the known constraints and the mass balance expressed by Eq. (1).

Since there are many sources, the mass-balance equation by itself is insufficient to constrain how much methane comes from each source. Nonetheless, we can use it to estimate the total annual emissions of methane. Based on a calculated lifetime of 10 years (τ), to be discussed later, and a global burden of 4800 Tg (C) obtained from global measurements (Figs. 1 to 5), and a current rate of change of about 20 Tg/yr (dC/dt; Fig. 1), we see from Eq. (1) that the total worldwide emissions from all sources should be about 500 Tg/yr. This is a useful benchmark.

There are a number of ways to improve on Eq. (1) that will allow us to reduce the uncertainties in the estimate of emissions from individual sources or combinations of sources. We will discuss three approaches here that have been useful in developing better global budgets. One method is to consider the long time series, such as the ice core data over several centuries, and apply Eq. (1) to two different time periods. This method uses the observed changes over long time periods to determine the ratio of anthropogenic to natural emissions. If we assume that several hundred years ago the concentration of methane in the atmosphere was determined entirely by natural processes, we can then estimate the emissions that would be required to satisfy

Eq. (1). This is particularly simple since at that time there are no significant trends, so $S = C/\tau$. Based on the ice core data for C, we can estimate the emissions to be $1700\,\text{Tg/yr}/10\,\text{yr} = 170\,\text{Tg/yr}$. We have already done the calculation for recent times suggesting present emissions of 500 Tg/yr. This would imply that there are new sources, presumably due to human activities, amounting to some 330 Tg/yr (Khalil and Rasmussen, 1990b). We have assumed that the lifetime of methane is the same now as it was a century or more ago. There is reason to believe that this is a good approximation, but the matter is open to question.

Another approach is to consider the latitudinal distribution of the various sources, determined by independent data or measurements. Then, a more detailed version of the mass-balance model, which takes into account the budget of methane over small regions of Earth's surface, can be used to determine whether the estimated rate of emissions from the sources is compatible with the measured atmospheric concentrations within each location. This method uses the latitudinal distribution to constrain the strength of the sources (Fung et al., 1991; Brown, 1993; Hein et al., 1997). For instance, we can rule out the oceans as the major source because that would require a more even distribution of methane across the hemispheres than is seen in Figure 4.

A recent approach at constraining the estimates of emissions uses carbon isotopes in methane. Normal measurements of methane cannot distinguish between the molecules of methane that come from one source or another, so only the total amount or concentration C is measured. It has become possible to measure the methane with different isotopes of carbon—specifically $^{12}CH_4$, $^{13}CH_4$, and $^{14}CH_4$ (Tyler, 1986; Stevens and Engelkemeir, 1988; Wahlen et al., 1989; Quay et al., 1991, Lassey et al., 1993). In this case, we can get more information on the global sources (and sinks, or loss processes) of methane since we can now have three equations similar to Eq. (1), one for each isotope. We now have to independently balance three types of methane ($^{12}CH_4$, $^{13}CH_4$, and $^{14}CH_4$) instead of just the sum of all types, using the same sources. There are fewer combinations of sources and emission rates that would balance all types than there are for just the total methane. Recently, stable isotopes of hydrogen in methane ($^{12}CH_3D$) have also been measured (Bergamaschi and Harris, 1995), which would add a fourth type of methane to constrain the sources. This work requires a knowledge of not only the isotopic combination of methane in the atmosphere, but also of the amounts of each type emitted by the sources, and the atmospheric lifetimes of each type of methane. Isotopic measurements hold considerable promise for reducing the uncertainties in the budget of methane.

4 SOURCES AND SINKS

Usually the global emission rate from a source is estimated by using a measured emission factor (grams of CH_4 emitted/day/unit source) and multiplying it by the number of such units in the world (units of source) and the time of year when emissions take place (days/year), which results in the grams of CH_4/year emitted by the source. The complexity of the estimate varies depending on the information

available. Once the budget is assembled, it must comply with the constraints discussed earlier.

Over the years many budgets have been proposed. Most of them were not entirely independent of previous estimates but tended to improve the estimates of emissions for one source or another. Two recent budgets are shown in Table 1. Both are "consensus"-type budgets in which several types of estimates by different researchers are put together. The first is from a NATO-sponsored Advanced Research Workshop (Khalil and Shearer, 1993). One of the goals of this project was to improve the budget based on direct measurements of emission factors and data on their global extrapolation. The second budget is from an assessment of the Intergovernmental Panel on Climate Change (IPCC) (Prather et al., 1995). The two budgets show one measure of the level of uncertainty that currently exists in the estimates of emissions from individual sources. Both these budgets are generally consistent with the known constraints, including the total emissions of around 500 Tg/yr discussed earlier. The budgets satisfy the constraints of the ratio of natural to anthropogenic emissions required by the ice core data. The budgets also agree on the major sources.

These budgets, like earlier ones, show that there are a few major sources. The major natural source is the wetlands, as has been known for a long time, since

TABLE 1 Comparison of Two Recent Budgets of Methane Sources

Source	NATO–ARW (1993)	IPCC (1994)
Natural sources (Tg)	Tg	Tg
Wetlands	110	115 (55–150)
Termites	20 (15–35)[a]	20 (10–50)
Open ocean	4	10 (5–50)
Marine sediments	(8–65)	
Geological	10 (1–13)	
Wild fire	2 (2–5)	
Other		15 (10–40)
Natural total	150	160 (110–210)
Anthropogenic source (Tg)	Tg	Tg
Rice agriculture	65 (55–90)	60 (20–100)
Animals	79	85 (65–100)
Manure	15	25 (20–30)
Landfills	22 (11–33)	40 (20–70)
Wastewater treatment	25 (12–38)	25 (15–80)
Biomass burning	50	40 (20–80)
Coal mining	46	30 (15–45)
Natural gas	30 (25–50)	40 (25–50)
Other anthropogenic	13 (7–30)	15 (5–30)
Low-temperature fuels	17	? (1–30)
Anthropogenic total	360	375 (300–450)
Total	510	535 (410–660)

[a] Numbers in parentheses show estimated range of source values.

methane has been called "marsh gas." Other natural sources are generally small but not well constrained. These include termites, oceans, and lakes. Most of the current sources are "anthropogenic." While these emissions are not directly from stacks and other easily identifiable icons of man-made pollution, they are a result of human activities nonetheless. These sources may be classified mostly as agricultural and from use of energy. Of these, rice agriculture, cattle, waste management, biomass burning, coal mining, and use of natural gas are the largest contributors. There are some moderate sized sources of a few teragrams/year that include transportation and fossil fuel combustion. There are perhaps many small sources that together fall within the range of uncertainty of the global emission rate and are therefore not included.

Methane is removed from the atmosphere by a number of processes. The most effective is reaction with hydroxyl radicals, or OH. The main process by which hydroxyl radicals are formed in the atmosphere occurs when sunlight splits an ozone molecule into O_2 and $O(^1D)$, an excited state of the oxygen atom. A few of these $O(^1D)$ atoms react with water vapor (H_2O) to form two OH radicals. OH has a lifetime of a few seconds and is removed mostly by its reaction with methane and CO. In addition to these major processes there are others that contribute to both the formation and destruction of OH radicals in the atmosphere (Thompson, 1992; DeMore et al., 1997). OH radicals are also responsible for removing many other gases from the atmosphere both man-made and natural. For example, many of the recently introduced chemicals (hydrofluorocarbons and hydrochlorofluorocarbons) that replace the chlorofluorocarbons are removed by OH radicals in the lower atmosphere. Methane is not only removed by OH, but there is enough methane in the atmosphere to control the abundance of OH and hence the oxidizing capacity of the atmosphere. Current estimates using photochemical models or proxy data suggest that the average concentration of OH is about 10^6 molecules/cm^3. At any location, OH concentrations vary greatly depending on latitude, altitude, season, and time of day. The rate constant (K) for the reaction of methane with OH is about 2.4×10^{-15} cm^3/molecule/sec 256 K (the average temperature of the atmosphere). Then, according to Eq. (1) the total loss of methane due to reactions with OH should be $C/\tau_{OH} = K_{[OH]}C$ or 400 Tg/yr after appropriate unit conversions. This corresponds to a lifetime of about 12 years due to reactions with OH alone.

Methane is also removed at Earth's surface by deposition and transport into the soils and then utilized by biological processes. This sink is estimated to be about 25 to 30 Tg/yr based on experimental field data. In the stratosphere methane is removed again by reacting with OH and also by other photolytic processes (DeMore et al., 1997). Recently, it has been suggested that there may be significant concentrations of Cl atoms in the marine boundary layer produced by precursors from the oceans. The concentration of these radicals is not known at present, but various estimates put it between 10^3 to 10^6 molecules/cm^3 (Singh and Kasting, 1988; Keene et al., 1990; Graedel and Keene, 1995, and references therein). At the upper limit this would constitute a significant sink for methane since it reacts 17 times faster with Cl atoms than with OH at the temperature in the boundary layer. Depending on how much of the marine atmosphere contains Cl radicals, this sink could be as large as 50 Tg/yr. Although we have not stated the sizable range of uncertainties in the estimates of

these smaller sinks, it should be noted that these calculations provide a composite lifetime of about 50 years, which when combined with the lifetime due to OH reactions results in a total global lifetime of about 10 years. This was used to impose the first constraint discussed earlier based on Eq. (1) whereby we calculated the total emissions to be about 500 Tg/yr.

5 PAST AND PRESENT TRENDS

While the budgets discussed so far represent current conditions, we need to know how these emissions have changed over the years before we can match the observed trends of concentrations shown in Figure 1 and represented in Eq. (1). Estimating past emissions, or a time series for each of these sources, is even more difficult than estimating current emission rates. The simplest approach is to assume that the current estimate of the total anthropogenic emissions is proportional to the human population, and the natural emissions have remained the same over the last 100 to 200 years. With these assumptions we can generate the emissions in Eq. (1) for the last century and calculate the expected concentrations. Although these assumptions are rough approximations, the results of this calculation explain the data quite well (Khalil and Rasmussen, 1994).

The assumption that anthropogenic emissions are proportional to human population breaks down in the recent decades. The atmospheric concentrations are not increasing as rapidly as would be expected if the anthropogenic emissions kept pace with the rising population. When we look at the data on the anthropogenic sources such as cattle populations or the area of rice fields, we find that these are not increasing any longer or are increasing very slowly. In the past these sources had been increasing at a rate proportional to human population. It seems then that there is a decoupling of the anthropogenic emissions from the human population. This circumstance makes the use of potential growth of human population an untenable surrogate for future emissions, even though it works well for the past.

An alternate and more detailed approach is to use the available agricultural and energy data to estimate how these emissions may have changed. Fortunately there are good records, going back a hundred years, on the number of cattle in the world and the hectares of rice harvested each year. Similar, but possibly less accurate estimates can also be made for the other anthropogenic sources based on archived records. We estimated the global emissions from the major sources over the last 100 years and calculated the expected concentrations using Eq. (1). The results are shown in Figure 6.

These results show that the available data for the calculation of global emissions over the last 100 years are in fact consistent with the observed concentrations shown in Figure 1. The long-term trends are driven by increases in rice agriculture and domestic cattle and collectively by the other anthropogenic sources. These same sources that led to the major increases of concentration over the last century are now stabilizing and causing the decreasing trends, at least in this model, and a more

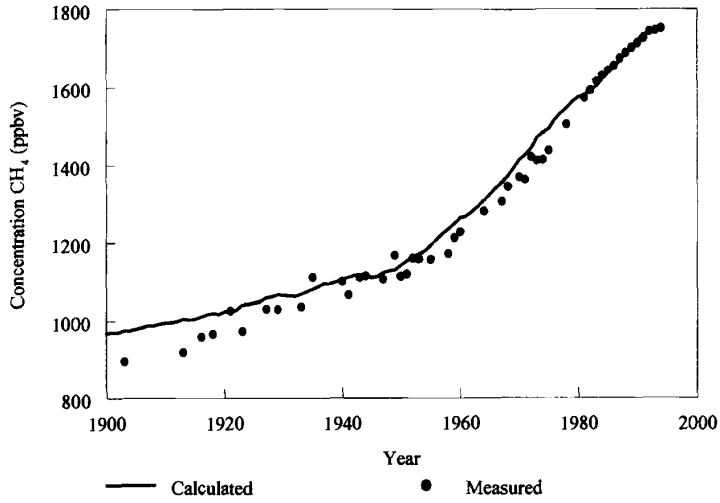


Figure 6 Comparison of the globally averaged calculated concentrations (shown by a smoothed line) with the measured concentrations of methane.

stable concentration of methane at present levels of 1750 ppbv. The results shown in Figure 6 are expanded for the recent decades and the trend both observed and calculated is plotted in Figure 7 (Khalil et al., 1996). These calculations lend support to the idea that the recent slowdown in the trend is caused by a stabilization of the major anthropogenic sources, mainly rice agriculture and domestic cattle.

The trends of methane can also be explained by changing levels of OH. If OH decreases, methane would increase and if it increases methane would decrease. Both these mechanisms have sometimes been discussed as alternatives to the explanation based on changing emission rates (Crutzen and Zimmermann, 1991; Thompson et al., 1993). It is more appropriate to consider this aspect as a contributing factor rather than an alternative explanation. For the long term, it is thought that OH may have decreased, adding to the trend of methane. This decrease of OH may have come about because both CO and CH_4 have increased over the last century, thus increasing the speed of removal of OH leading to lower concentrations. This process may have been partially compensated by increased ozone that can lead to greater production and by other chemical feedbacks. Calculations suggest that the long-term changes of OH are probably small and not sufficient to explain the major part of the observed increase (Khalil and Rasmussen, 1993; Pinto and Khalil, 1991; Lu and Khalil, 1991).

In more recent decades, there has been evidence for the depletion of the ozone layer due to the man-made chlorofluorocarbons. This would cause an increase of ultraviolet (UV) radiation in the troposphere where it would stimulate the splitting of

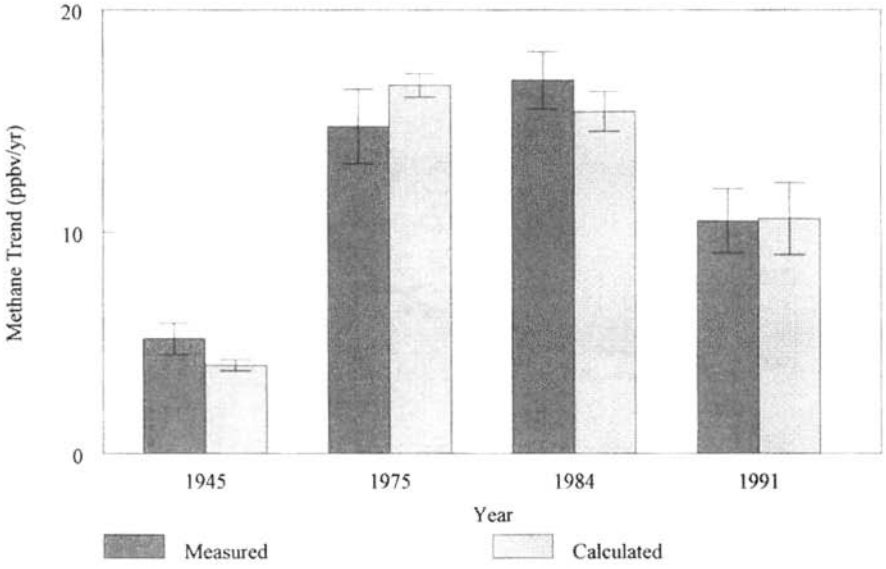


Figure 7 Comparison of trends calculated from measured methane concentrations vs. modeled methane concentrations.

the O_3 molecules discussed earlier into O_2 and $O(^1D)$, thus increasing the production of OH (Madronich and Granier, 1992). The increased OH would cause the trend of methane to slow down. Such a mechanism may contribute to the slowdown of the trend, but there is no experimental evidence that can pin down the magnitude of the OH trend. The increase of OH, if it is occurring, appears to be small and not sufficient to explain the entire observed slowdown (Krol et al., 1998). It will take more work before we can say how much of the current trend is affected by possible changes of OH and how much is from the slowdown of emissions. For the present, however, it seems that the slowdown of emissions can be estimated, and these estimates of changing emissions are sufficient to explain the general pattern of the observations as shown in Figure 7.

6 DISCUSSION AND COMMENTARY

The state of knowledge about the methane cycle is that we have a clear understanding of the global distributions and trends in recent decades and over the last century for which ice core data are used. This is a directly measurable component of the global balance. There are no substantive differences among the various groups who have measured methane in the atmosphere. Field studies of emissions from the various sources are also in broad agreement, and the differences that have been observed are explained by environmental variables. Extrapolation of the field data

to global emission rates remains a major source of uncertainty, leaving a sizable uncertainty in the estimates of global emissions from each source. The main features of the trends, both the increases over the last century and the slowdown of the trend in recent times, are consistent with what we know about the change of emissions from anthropogenic sources. There is enough uncertainty that trends caused by changes of OH concentrations can be accommodated.

Although the current understanding of the methane distribution and trends can be explained by the known sources and sinks, the very nature of these explanations clouds our ability to predict future concentrations. We see that the major anthropogenic sources—rice fields, cattle and also biomass burning—are all stabilizing not because of legislated controls, but because there are natural limitations to the growth of these sources. These sources will not keep pace with increasing population as new technologies make it unnecessary to do so. For instance, new high yielding varieties of rice do not require as much land or time in the growing season to produce the same amount of rice as before, thus reducing the emissions of methane per bushel of rice grown. If the anthropogenic sources could be related to population in the future, it would then be easier to predict future emissions under various assumptions of population growth—but this is not possible as we have discussed. Various scenarios that had been hypothesized are no longer likely (Alcamo et al., 1995). Past estimates of the doubling of methane to 3 to 4 ppmv are now unlikely with no known sources that could increase sufficiently to cause such high concentrations. Perhaps the only prediction that can be made is that it is quite unlikely that the concentrations of methane will increase substantially or double in the next decade or two. This is good news for global warming since it is not likely to be as much as previously expected from the increase of methane. As such, methane will continue to play an important role in the global environment, but this role is not likely to increase for years to come.

REFERENCES

Alcamo, J., A. Bouwman, J. Edmonds, A. Grübler, T. Morita, and A. Sugandhy, An evaluation of the IPCC IS92 emission scenarios, in J. T. Houghton, L. G. Meira Filho, J. Bruce, H. Lee, B. A. Callander, E. Haites, N. Harris, and K. Maskell (Eds.), *Climate Change 1994, Radiative Forcing of Climate Change and an Evaluation of the IPCC IS92 Emission Scenarios*, Intergovernmental Panel on Climate Change, Cambridge University Press, Great Britain, 1995.

Bergamaschi, P., and G. W. Harris, Measurement of stable isotope ratios ($^{13}CH_4/^{12}CH_4$; $^{12}CH_3D/^{12}CH_4$) in landfill methane using a tunable diode laser absorption spectrometer, *Global Biogeochem. Cycles*, *9*, 439–447, 1995.

Blake, D. R., and F. S. Rowland, Continuing worldwide increase in tropospheric methane, 1978 to 1987, *Science*, *239*, 1129–1131, 1988.

Brown, M., Deduction of emissions of source gases using an objective inversion algorithm and a chemical transport model, *J. Geophys. Res.*, *98*, 12639–12660, 1993.

Chappellaz, J., J. M. Barnola, D. Raynaud, Y. S. Korotkevich, and C. Lorius, Ice-core record of atmospheric methane over the past 160,000 years, *Nature*, *345*, 127–131, 1990.

Crutzen, P. J., and P. H. Zimmermann, The changing photochemistry of the troposphere, *Tellus*, *43AB*, 136–151, 1991.

DeMore, W. B., S. P. Sander, C. J. Howard, A. R. Ravishankara, D. M. Golden, C. E. Kolb, R. F. Hampson, M. J. Kurylo, and M. J. Molina, *Chemical Kinetics and Photochemical Data for Use in Stratospheric Modeling*, JPL Publication 97-4, National Aeronautics and Space Administration Jet Propulsion Laboratory, Pasadena, CA, 1997.

Dlugokencky, E. J., L. P. Steele, P. M. Lang, and K. A. Masarie, The growth rate and distribution of atmospheric methane, *J. Geophys. Res.*, *99*, 17021–17043, 1994.

Etheridge, D. M., G. I. Pearman, and P. J. Fraser, Changes in tropospheric methane between 1841 and 1978 from a high accumulation-rate Antarctic ice core, *Tellus*, *44B*, 282–294, 1992.

Fabian, P., R. Borchers, G. Glentje, W. A. Matthews, W. Seiler, H. Giehl, K. Bunse, F. Müller, U. Schmidt, A. Volz, A. Khedim, and F. J. Johnen, The vertical distribution of stable trace gases at mid-latitudes, *J. Geophys. Res.*, *86*, 5179–5184, 1981.

Fung, I., J. John, J. Lerner, E. Matthews, M. Prather, L. P. Steele, and P. J. Fraser, Three-dimensional model synthesis of the global methane cycle, *J. Geophys. Res.*, *96*, 13033–13065, 1991.

Graedel, T. E., and W. C. Keene, Tropospheric budget of reactive chlorine, *Global Biogeochem. Cycles*, *9*, 47–77, 1995.

Hein, R., P. J. Crutzen, and M. Heimann, An inverse modeling approach to investigate the global atmospheric methane cycle, *Global Biogeochem. Cycles*, *11*, 43–76, 1997.

Keene, W. C., A. A. P. Pszenny, D. J. Jacob, R. A. Duce, J. N. Galloway, J. J. Schultz-Tokos, H. Sievering, and J. F. Boatman, The geochemical cycling of reactive chlorine through the marine troposphere, *Global Biogeochem. Cycles*, *4*, 407–430, 1990.

Khalil, M. A. K., and R. A. Rasmussen, Sources, sinks, and seasonal cycles of atmospheric methane, *J. Geophys. Res.*, *88*, 5131–5144, 1983.

Khalil, M. A. K., and R. A. Rasmussen, Atmospheric methane: Recent global trends, *Environ. Sci. Technol.*, *24*, 549–553, 1990a.

Khalil, M. A. K., and R. A. Rasmussen, Constraints on the global sources of methane and an analysis of recent budgets, *Tellus*, *42B*, 229–236, 1990b.

Khalil, M. A. K., and R. A. Rasmussen, Decreasing trend of methane: Unpredictability of future concentrations, *Chemosphere*, *26*, 595–608, 1993.

Khalil, M. A. K., and M. J. Shearer, Sources of methane: An overview, in M. A. K. Khalil (Ed.), *Atmospheric Methane: Sources, Sinks, and Role in Global Change*, NATO ASI Series I: *Global Environmental Change*, Vol. 13, Springer-Verlag, Berlin, 1993.

Khalil, M. A. K., and R. A. Rasmussen, Global emissions of methane during the last several centuries, *Chemosphere*, *29*, 833–842, 1994.

Khalil, M. A. K., R. A. Rasmussen, and F. Moraes, Atmospheric methane at Cape Meares: Analysis of a high resolution data base and its environmental implications, *J. Geophys. Res.*, *98*, 14753–14770, 1993.

Khalil, M. A. K., R. A. Rasmussen, and M. J. Shearer, Trends of atmospheric methane during the 1960s and 1970s, *J. Geophys. Res.*, *94*, 18279–18288, 1989.

Khalil, M. A. K., M. J. Shearer, and R. A. Rasmussen, Atmospheric methane over the last century, *World Resource Rev.*, *8*, 481–492, 1996.

Krol, M., P. J. van Leeuwen, and J. Lelieveld, Global OH trend inferred from methylchloroform measurements, *J. Geophys. Res.*, *103*, 10697–10711, 1998.

Lassey, K. R., D. C. Lowe, C. A. M. Brenninkmeijer, and A. J. Gomez, Atmospheric methane and its carbon isotopes in the Southern Hemisphere: Their time series and an instructive model, *Chemosphere*, *26*, 95–109, 1993.

Lu, Y., and M. A. K. Khalil, Tropospheric OH: Model calculations of spatial, temporal, and secular variations, *Chemosphere*, *23*, 397–444, 1991.

Madronich, S., and C. Granier, Impact of recent total ozone changes on tropospheric ozone photodissociation, hydroxyl radicals, and methane trends, *Geophys Res. Lett.*, *19*, 465–467, 1992.

Pinto, J. P., and M. A. K. Khalil, The stability of tropospheric OH during ice ages, inter-glacial epochs and modern times, *Tellus*, *43B*, 347–352, 1991.

Prather, M., R. Derwent, D. Ehhalt, P. Fraser, E. Sanhueza, and X. Zhou, Other trace gases and atmospheric chemistry, in J. T. Houghton, L. G. Meira Filho, J. Bruce, H. Lee, B. A. Callander, E. Haites, N. Harris, and K. Maskell (Eds.), *Climate Change 1994, Radiative Forcing of Climate Change and an Evaluation of the IPCC IS92 Emission Scenarios*, Intergovernmental Panel on Climate Change, Cambridge University Press, Great Britain, 1995.

Quay, P. D., S. L. King, J. Stutsman, D. O. Wilbur, L. P. Steele, I. Fung, R. H. Gammon, T. A. Brown, G. W. Farwell, P. M Grootes, and F. H. Schmidt, Carbon isotopic composition of atmospheric CH_4: Fossil and biomass burning source strengths, *Global Biogeochem. Cycles*, *5*, 25–47, 1991.

Rasmussen, R. A., and M. A. K. Khalil, Increase in the concentration of atmospheric methane, *Atmos. Environ.*, *15*, 883–886, 1981.

Rasmussen, R. A., and M. A. K. Khalil, Atmospheric methane in the recent and ancient atmospheres: Concentrations, trends, and interhemispheric gradient, *J. Geophys. Res.*, *89*, 11599–11605, 1984.

Schmidt, U., A. Khedim, D. Knapsa, G. Kulessa, and F. J. Johnen, Stratospheric trace gas distributions observed in different seasons, *Adv. Space Res.*, *4*, 131–134, 1984.

Schmidt, U., G. Kulessa, E. Klein, E.-P. Röth, P. Fabian, and R. Borchers, Intercomparison of balloon-borne cryogenic whole air samplers during the MAP/GLOBUS 1983 campaign, *Planet. Space Sci.*, *35*, 647–656, 1987.

Singh, H. B., and J. F. Kasting, Chlorine-hydrocarbon photochemistry in the marine troposphere and lower stratosphere, *J. Atmos. Chem.*, *7*, 261–285, 1988.

Steele, L. P., E. J. Dlugokencky, P. M. Lang, P. P. Tans, R. C. Martin, and K. A. Masarie, Slowing down of the global accumulation of atmospheric methane during the 1980's, *Nature*, *358*, 313–316, 1992.

Stevens, C. M., and A. Engelkemeir, Stable carbon isotope composition of methane from some natural and anthropogenic sources, *J. Geophys. Res.*, *93*, 725–733, 1988.

Taylor, F. W., A. Dudhia, and C. D. Rogers, Proposed reference models for nitrous oxide and methane in the middle atmosphere, in G. M. Keating (Ed.), *Handbook for MAP*, Vol. 31, 1989, pp. 67–79. Middle Atmosphere Program ISCU SCOTEP, U. of Illinois, Urbana , IL, USA.

Thompson, A. M., The oxidizing capacity of the Earth's atmosphere: Probable past and future changes, *Science*, *256*, 1157–1165, 1992.

Thompson, A. M., J. A. Chappellaz, and I. Y. Fung, The atmospheric CH_4 increase since the Last Glacial Maximum (2) Interactions with oxidants, *Tellus*, *45B*, 242–257, 1993.

Tyler, S. C., Stable carbon isotope ratios in atmospheric methane and some of its sources, *J. Geophys. Res.*, *91*, 13232–13238, 1986.

Wahlen, M., N. Tanaka, R. Henry, B. Deck, J. Zeglen, J. S. Vogel, J. Southon, A. Shemesh, R. Fairbanks, and W. Broecker, Carbon-14 in methane sources and in atmospheric methane: The contribution from fossil carbon, *Science*, *245*, 286–290, 1989.

CHAPTER 7

BIOGENIC NON-METHANE HYDROCARBONS

MARCY E. LITVAK

1 INTRODUCTION

Nonmethane volatile organic compounds (NMVOCs) are emitted from a wide variety of both anthropogenic and biogenic sources. Major anthropogenic sources of NMVOCs include combustion of fossil fuels, solvent evaporation and biomass burning, while direct emissions from plants are the largest biogenic source. Over 90% of the total NMVOCs entering the atmosphere are biogenic (Guenther et al., 1995; Müller, 1992). Recent estimates of the upper limit of global NMVOC emissions from biogenic sources range from 1000 to 1500 Tg C/yr ($1\,\text{Tg} = 10^{12}\,\text{g}$), an amount equivalent to the total methane flux from both biogenic and anthropogenic sources (Guenther et al., 1995).

In the atmosphere, NMVOCs are typically very reactive (lifetimes range from minutes to days) and play significant roles in many aspects of atmospheric chemistry. NMVOCs are a key component of the photochemical processes that form ozone and other secondary products in the planetary boundary layer (Fehsenfeld et al., 1992). The other products produced include organic acids, organic nitrates, aerosols, acetone, formaldehyde, and carbon monoxide (Kasting and Singh, 1986; Trainer et al., 1987; Chameides et al., 1988; Jacob and Wofsy, 1988; Andreae et al., 1988; Fehsenfeld et al., 1992). These products are relevant in that they can contribute to both air pollution and climate change. Ozone is not only a potent greenhouse gas but can impact human health and plant productivity. Organic nitrates such as PAN (peroxyacetyl nitrate) are phytotoxic, an important component of urban smog, and also provide a mechanism for transporting reactive nitrogen (NO and NO_2, together referred to as NO_x) over large distances (Sillman and Samson, 1995). Organic

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

aerosol particles scatter light at all visible wavelengths, which creates haze and decreases visibility (Andreae and Crutzen, 1997; Pandis et al., 1991). Finally, NMVOCs and carbon monoxide are considerably more reactive toward the hydroxyl radical (OH) than is methane. Increased levels of CO and NMVOCs therefore can significantly suppress OH concentrations and thus the oxidative capacity of the troposphere, resulting in a longer atmospheric lifetime for methane (Brasseur and Chatfield, 1991; Fehsenfeld et al., 1992).

In many localized areas, biogenic hydrocarbons play a dominant role in generating tropospheric ozone locally and in rural areas downwind from these urban centers. In Atlanta, Georgia, which has a high density of NMVOC-emitting plant species, Geron et al. (1995) estimated that with current NO_x and biogenic hydrocarbon emissions, even if anthropogenic hydrocarbon emissions were reduced to zero, ozone levels would be above the National Ambient Air Quality Standard (NAAQS). In rural areas long distances from polluted plumes, anthropogenic VOCs are so diluted that isoprene and terpenes emitted from vegetation alone are enough to sustain ozone production (Roselle et al., 1991; Hagerman et al., 1997).

The list of NMVOCs emitted from biogenic sources includes well over 1000 compounds. In many ecosystems, isoprene and monoterpenes are the predominant biogenic hydrocarbons emitted, and these compounds account for over half of the total global NMVOC fluxes from biogenic sources (Table 1). However, recent

TABLE 1 Major Biogenic Methane and NMVOC Sources, Source Strength, and Atmospheric Lifetimes[a]

VOC	Primary Natural Sources	Estimated Annual Global Emission (Tg C)	Reactivity in Atmosphere (lifetime in days)
Methane	Wetlands, rice paddies	319–412	4000
Isoprene	Plants	175–503	0.2
Monoterpenes	Plants	127–480	0.1–0.2
Ethene	Plants, soils, oceans	8–25	1.9
Other reactive VOCs (e.g., acetaldehyde, formaldehyde MBO, hexenal family)	Plants	~260	<1
Other less reactive VOCs (e.g., methanol, ethanol, formic acid, acetic acid, acetone)	Plants, soils	~260	>1

[a]Adapted from Fall (1999). Data are derived from Singh and Zimmerman (1992), Conrad (1995), Guenther et al. (1995), Andreae and Crutzen (1997), and Rudolph (1997).

studies have emerged indicating that many other hydrocarbons, particularly oxygenated VOCs, also provide a significant contribution to the total biogenic NMVOC flux (Isidorov et al., 1985; Arey et al., 1991; Winer et al., 1992; König et al., 1995; Helmig et al., 1999). Quantitative measurements of most of these "other" NMVOCs are scarce because of the wide variety of sources and difficulty in reliable identification and quantification of these compounds.

In this chapter, the major classes of hydrocarbons and their oxygenated derivatives emitted to the atmosphere from biogenic sources are reviewed. Information is also provided on ambient mixing ratios, regional and global distribution, and the primary controlling factors over emissions of these classes of biogenic NMVOC's. Of the large group of biogenic NMVOCs that are highly reactive in the atmosphere (have lifetimes of less than one day), this chapter will mainly focus on isoprene, monoterpenes, ethene, propene, butene, acetaldehyde, formaldehyde, 2-methyl-3-buten-2-ol (MBO), and the hexenal family compounds (hexenylacetate, 2-hexenal, 3-hexenol, and hexanal). The nonreactive biogenic NMVOCs (lifetimes of more than one day) covered here include methanol and ethanol, acetone, ethane, and acetic and formic acid. Emissions of alkanes (e.g. ethane, propane, butane) from terrestrial and oceanic natural sources are very low (Lindskog, 1997; Guenther et al., 1994) and are not covered here.

2 BIOGENIC NMVOCs

Isoprene

Isoprene (2-methyl-1,3-butadiene) was first recognized as an emission from plant tissues in the late 1950s (Sanadze, 1991) (Fig. 1). Until recently, it was thought that isoprene was synthesized by the mevalonic acid pathway. It is now known that isoprene is produced in chloroplasts by the glyceraldehyde-3-phosphate pathway, in both an enzyme-dependent (catalyzed by the enzyme isoprene synthase) and nonenzymatic manner (Lichtenthaler et al., 1997). Release to the atmosphere is instantaneous following synthesis, and is the result of simple diffusion of isoprene through cell membranes into the intercellular air spaces and out of pores on the leaf surface, called stomata.

Isoprene emission rates vary among species from 0.1 to 70 $\mu g/g\,dw\,h$. Not all plants have the ability to produce and emit significant amounts of isoprene. A compilation of species-level isoprene emission screenings from over 800 species of higher plants indicate that, in general, most isoprene emitters are woody deciduous species, although some ferns, vines, and other herbaceous species also emit significant amounts of isoprene (Harley et al., 1999). Phylogenetic patterns are hard to find, as many plant families that contain isoprene emitters, contain nonemitters as well. High isoprene emitting species have been found in the genera *Quercus* (oaks), *Populus* (aspen and poplars), and *Liquidambar* (sweetgum).

Leaf temperature and light intensity are the primary environmental controllers of short-term (hours to days) changes in isoprene production and emission rates from plant foliage (Guenther et al., 1993; Sharkey et al., 1999). Isoprene emissions show

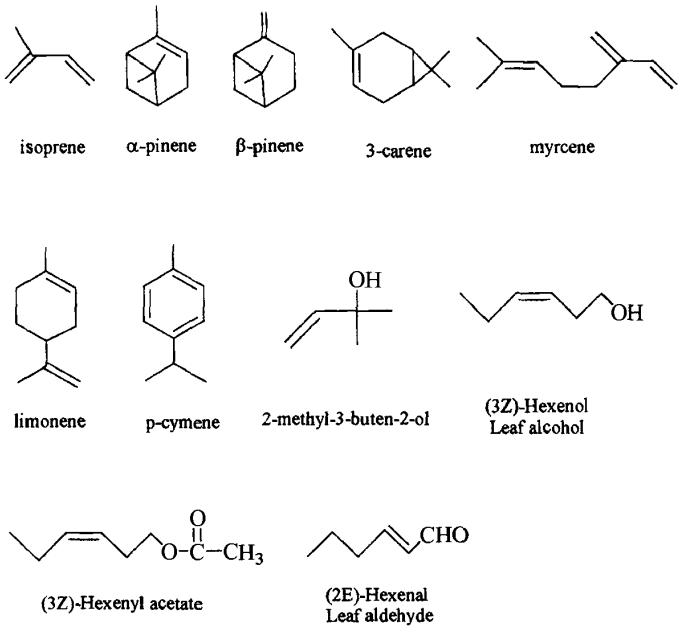


Figure 1 Selected nonmethane hydrocarbons emitted from natural sources.

typical Arrhenius temperature kinetics with species-dependent temperature optima that range from 36 to 40°C (Guenther et al., 1993). Long-term factors that influence isoprene emission rates include light and temperature conditions in which leaves develop, water and nutrient availability, and disease (Monson et al., 1994; Lerdau and Throop, 2000; Anderson et al., 2000; Harley et al., 1994).

Although plants may lose a significant fraction of fixed carbon to isoprene production, it is not known if the production and emission of isoprene serves an adaptive role in plant tissues. One hypothesis is that isoprene protects photosynthetic apparatus against damage from exposure to high temperature and light intensity (Sharkey, 1997). Another possibility is that isoprene scavenges reactive oxidants inside the leaf that can damage plant tissues (Harley et al., 1999).

Over 90% of total isoprene fluxes are from canopy foliage (Guenther, 1999). Emissions from bacteria and fungi in soils and ground cover foliage of mosses and ferns make up the bulk of the remaining natural source strength (9%). Mammals and marine algae and anthropogenic sources (automobile emissions and industrial processes) each contribute less 1% of the global isoprene flux. Although oxidation in the atmosphere is the primary sink for isoprene, microbial consumption in soils is a small net sink as well (Cleveland and Yavitt, 1998).

Typical surface layer mixing ratios of isoprene in the summer range from less than 1 ppbv in a Colorado pine forest (Goldan et al., 1993) to 8 ppbv in the tropical

rain forest (Rasmussen and Khalil, 1988) and can be as high as 20 ppbv in rural forests in the southeastern United States (Hagerman et al., 1997). Isoprene ambient concentrations are highest in the summer months and typically show strong diurnal patterns where concentrations sharply increase after sunrise to a maximum in the afternoon and fall to zero at night (Fehsenfeld et al., 1992). This pattern can be explained by the dependence of isoprene emission rates on both temperature and light.

Factors that influence ambient isoprene mixing ratios include emission rates, season, stability of the atmosphere, origin of air masses, and oxidation capacity of the atmosphere (Steinbrecher, 1997). Mixing ratios typically decrease rapidly with altitude since isoprene reacts rapidly in the troposphere with both OH and ozone (e.g., Helmig et al., 1998). Some isoprene has been found in the free troposphere, but only in very low and variable amounts.

Monoterpenes

Monoterpenes ($C_{10}H_{16}$) are a class of structurally diverse compounds produced by over 46 families of flowering plants (e.g., mint, composite, and citrus families), almost all conifers, and some species of liverworts (Banthorpe and Charlwood, 1980; Adam et al., 1996). The accumulation and emission of these compounds directly defends plant tissues against herbivores and pathogens, indirectly defends plant tissues by attracting predators of herbivores, and attracts floral pollinators [reviewed in Langenheim (1994)]. The array of over 1000 different monoterpene structures includes acyclic, monocyclic, and bicyclic forms that can be simple hydrocarbons (e.g., α-pinene, β-pinene, myrcene, δ-3-carene, limonene, p-cymene) or oxygenated derivatives (e.g., 1-8 cineole, linalool, camphor) (Fig. 1). The specific monoterpenes produced and emitted from each species is under tight genetic control, and typically only a few monoterpenes dominate the emissions profile of each species.

Monoterpenes, like isoprene, are synthesized in chloroplasts of specialized tissues by the glyceraldehyde-3-phosphate pathway (Lichtenthaler et al., 1997). Two 5C "isoprene" units condense to form a 10C precursor, which is transformed into the myriad of monoterpenes by cyclization reactions catalyzed by the enzymes monoterpene cyclases (Gershenzon and Croteau, 1991). Monoterpenes typically accumulate in storage structures in plant tissues such as glandular trichomes (mints), resin cysts and ducts (conifers), or cavities (eucalypts).

Release to the atmosphere of these stored pools is dependent upon both volatilization and diffusion processes. Plant foliage is the largest source of monoterpene emissions (over 90% of the total global flux) (Guenther, 1999). The remaining fluxes are from woody tissues, buds, cones, and flowers.

In conifers, total foliar monoterpene emission rates vary from 0.01 to 10 μg/g dw h. Emission of monoterpenes is a diffusive process controlled primarily by the influence of needle temperature on monoterpene vapor pressure and monoterpene concentration in the resin ducts and the diffusive resistance of the tissue to volatile losses (Tingey et al., 1991). Other controls over the emission rates of stored

pools include leaf age (Lerdau et al., 1997), phenology (Fukui and Doskey, 1998; Cao et al., 1997; Lerdau et al., 1995), herbivory (Litvak and Monson, 1998; Litvak et al., 1999), relative humidity (Dement et al., 1975), foliar moisture (Lamb et al., 1985), and water stress (Yani et al., 1993).

Atmospheric monoterpene concentrations in four rural sites in the southeastern United States ranged from 0.32 to 0.63 ppbv in the summer, and from 0.125 to 0.19 ppbv in the winter (Hagerman et al., 1997). Maximum summer ambient concentrations of total monoterpenes were 0.80 ppbv above a lodgepole pine forest in Colorado (Roberts et al., 1983) and 0.38 above a ponderosa pine plantation in the Sierra Nevada (Lamanna and Goldstein, 1999). Clear diurnal patterns in ambient concentrations of monoterpenes are not as pronounced as those observed for isoprene, but concentrations are often highest at night and lowest during the day (e.g., Lamanna and Goldstein, 1999; Hagerman et al., 1997). Both vertical mixing and chemical loss are important controllers of ambient monoterpene concentrations. In clean air, vertical mixing and dispersion are the most important factors (Hewitt et al., 1995). Because monoterpenes are still emitted at night when atmospheric conditions are relatively stable, concentrations often increase after sunset until the breakdown of these conditions in the morning.

Recent evidence also indicates some species (e.g., Holm oak *Quercus ilex*, Norway spruce *Picea abies*, *Pinus pinea*, and *Acer saccharinum*) produce and emit monoterpenes that do not accumulate in pools (Steinbrecher et al., 1993; Loreto et al., 1996; Staudt et al., 1997). These monoterpenes are emitted at relatively high rates in a light- and temperature-dependent manner very similar to that observed for isoprene and are sensitive to water stress (Bertin and Staudt, 1996), and phenology (Staudt et al., 1997). Many questions remain concerning the physiological and ecological controls over light-dependent production and emission of monoterpenes as well as the specific roles these compounds play in plant tissues.

The aromatic *p*-cymene (1-methyl-4-isopropyl-benzene) is the only volatile arene emitted from vegetation. Trace fluxes of *p*-cymene have been measured from conifers, sage, and eucalyptus but together are equivalent to only 1% of the estimated global monoterpene source strength (Fehsenfeld et al., 1992).

Light Alkenes

Substantial quantities of ethene, propene, and butenes, are released annually from automobiles, industry, and biomass burning (estimated at 10 Tg/yr). However, atmospheric measurements of alkenes made in remote areas that are not impacted by urban or industrial emissions suggest the presence of biogenic sources of light alkenes as well (Lamanna and Goldstein, 1999; Goldstein et al., 1996; Heikes et al., 1996a; Rudolph, 1997).

Emissions from terrestrial ecosystems, particularly plant tissues, make up the bulk of the total global emissions of ethene from natural sources (Sawada and Totsuka, 1986). Ethene functions as a hormone in plant tissues that triggers growth and developmental processes including seed germination, flowering, fruit ripening, senescence, and growth regulation [reviewed in Abeles et al. (1992)]. In

addition, ethene is a well-known stress indicator and may play a role in triggering plant defense mechanisms. The amino acid L-methionine is enzymatically converted to ethene in a two-step process involving the intermediate 1-aminocyclo-propane-1-carboxylate (ACC) [reviewed in Fall (1999)]. Production and emission rates of ethene vary with species, tissue type, and phenology and are significantly induced in response to wounding, air pollution, insect and pathogen attack, drought, water-logging, high and low temperatures, and gamma radiation [reviewed in Abeles et al. (1992)]. Global estimates of ethene fluxes from undisturbed canopy foliage are 2 to 4 Tg/yr (Table 1; Rudolph, 1997).

Ethene is also emitted in small quantities from soil microorganisms. Fluxes are correlated with the organic matter content in soil and on a global scale are 2.6 to 3.7 Tg/yr (Rudolph, 1997). Fluxes of ethene, propene, butene, and acetylene have been measured from wetlands but are insignificant on a global scale.

Due to the short lifetimes, measurement difficulty, and wide variety of sources of ethene and propene, ambient concentrations of these compounds are variable. Mean summertime emission rates of ethene, propene, and 1-butene from a deciduous forest in the northeastern United States were 2.6, 1.1, and 0.4×10^{10} molecules/cm^{-2} s, respectively (Goldstein et al., 1996). In this forest, biogenic emissions of propene and 1-butene exceeded the anthropogenic emissions, while biogenic emissions of ethene were equivalent to 50% of emissions from anthropogenic sources. Maximum ambient concentrations above the forest were 0.2, 0.95, and 0.08 ppbv for propene, ethene, and 1-butene, respectively (Goldstein et al., 1996). Lamanna and Goldstein (1999) also observed a local biogenic source for ethene and propene in measurements above a Sierra Nevada ponderosa pine plantation where ambient concentrations of these compounds varied between 0.18 and 0.45 ppbv.

Photochemical degradation of dissolved organic carbon (DOC) released by marine algae results in an estimated global emission rate of 5 Tg/yr for ethene, propene, butenes, and acetylene from ocean surface water [reviewed in Rudolph (1997)]. Fluxes vary seasonally, increase with DOC and light intensity (particularly shorter wavelengths), and depend strongly on DOC, biological activity of the algae, and wind speed (drives the exchange at the air–sea interface) (Ratte et al., 1995). Fluxes are inferred from a combination of atmospheric measurements, seawater measurements, air–sea exchange rates and photochemical models, and encompass large uncertainties. In the remote marine boundary layer and free troposphere over the South Atlantic and western Indian Oceans, ethene and propene concentrations were less than 20 and 6 ppt, respectively (Heikes et al., 1996a).

Alcohols

A C5 alcohol, 2-methyl-3-buten-2-ol (MBO), was recently identified in air samples taken in a Colorado pine forest in concentrations higher than isoprene (up to 3.5 ppbv; Goldan et al., 1993). It is now known that MBO is emitted at relatively high rates from many pine species that grow predominantly in the western United States (up to 70 μg C/g h). Fluxes of MBO, like isoprene, are both light and temperature dependent, suggesting that MBO is emitted immediately following production

rather than stored in specialized structures (Harley et al., 1998). Although the production mechanism of MBO in plant tissues is not well known, there is some evidence that it is derived from a 5C precursor of isoprene (Fall, 1999).

Most of the nonreactive other NMVOC flux in Table 1 is contributed by methanol (Guenther et al., 1995). Methanol fluxes measured from leaves are comparable to isoprene and monoterpenes and vary from 0.2 to 40 μg C/h g dry weight (MacDonald and Fall, 1993; Nemecek-Marshall et al., 1995). Emission rates of methanol are highest in young leaves and vary with phenology, leaf damage, and stomatal conductance (Nemecek-Marshall et al., 1995; Fukui and Doskey, 1998). Significant fluxes of methanol and ethanol have also been measured from decaying plant material. Warneke et al. (1999) estimate that globally, emissions from decaying plant material alone could account for 18 to 40 Tg of methanol per year.

Methanol was one of the most abundant VOCs detected above a pine forest canopy in the rural southeastern United States, with summertime mixing ratios of 10 to 20 ppbv (Goldan et al., 1995). Like isoprene, ambient methanol mixing ratios in these studies varied diurnally and peaked in the midafternoon, suggesting that at least in these rural forested areas, methanol was derived primarily from biogenic sources. At a rural site in Colorado, maximum summertime ambient mixing ratios were 6 ppbv (Goldan et al., 1997). Relatively high concentrations of methanol have been detected in the free troposphere, particularly in the northern midlatitudes (0.6 to 0.8 ppbv in northern areas and 0.4 ppbv in southern areas) (Singh et al., 1995). In addition to direct emissions from vegetation, sources of atmospheric methanol include fossil fuel use, biomass burning, and tropospheric production.

Other nonterpenoid alcohols emitted from many agricultural crops, grasses, pastures and forest trees include 3Z-hexenol (leaf alcohol), ethanol, methyl propanol, butanol, and octanol (Isidorov et al., 1985; Arey et al., 1993; Macdonald and Fall, 1993; König et al., 1995; Puxbaum, 1997; Kirstine et al., 1998; Helmig et al., 1999). Production of many of these alcohols, particularly leaf alcohol and ethanol, varies with phenology and is triggered by physical injury and environmental stress (Kirstine et al., 1998; MacDonald et al., 1989). Fluxes of alcohols and other oxygenated VOC's released during the process of crop harvesting may be large enough to have a short-term influence on local air quality (Karl et al., 2001).

Aldehydes and Ketones

Many aldehydes and ketones that are detected in the atmosphere, e.g., acetaldehyde (ethanal), formaldehyde, propanal, butanal, acetone, and butenone, have both anthropogenic and biogenic sources. The dominant sources of these species are fossil fuel combustion, biomass burning, and photochemical oxidation of man-made and natural hydrocarbons, but direct emissions from a variety of forest trees, shrubs, grasses, ferns and mosses occur as well (Isidorov et al., 1985; MacDonald and Fall, 1993; Kotzias et al., 1997; Fukui and Doskey, 1998).

Acetone in plant tissues is produced through fatty acid oxidation [reviewed in Fall (1999)]. Small acetone fluxes have been measured from live plant foliage, decaying

vegetation, and seeds and buds of many conifer species, suggesting at least some of the acetone measured in forest canopies and the free troposphere is contributed by natural sources (Fukui and Doskey, 1998; Warneke et al., 1999; Kotzias et al., 1997; MacDonald and Fall, 1993). Warneke et al. (1999) estimated that on a global scale, decaying vegetation emits 6 to 8 Tg of acetone annually.

Acetone is one of the most abundant oxygenated species in the remote atmosphere (Singh et al., 1995). In the free troposphere over the Pacific Ocean, Singh et al. (1995) measured acetone concentrations that range from 0.5 ppbv in the northern latitudes to 0.25 ppbv in the southern latitudes.

Singh et al. (1995) estimated that direct biogenic emissions account for 21% of the total global acetone source. Like methanol, acetone was one of the most abundant VOCs measured above several rural forested areas in Alabama (4 to 7 ppbv; Goldan et al., 1995). Summertime acetone mixing ratios in the Sierra Nevada above a ponderosa pine plantation ranged from 1.5 to 8 ppbv (Lamanna and Goldstein, 1999). At this site, biogenic sources (primarily oxidation of the alcohol MBO) accounted for 45% of the acetone concentrations that exceeded background levels (Goldstein and Schade, 1999). In remote and rural forested regions in Europe, ambient surface acetone concentrations varied between 0.2 and 2.2 ppbv (Solberg et al., 1996). Solberg et al. (1996) observed a strong seasonal dependence of acetone mixing ratios at these sites where summertime maximum acetone concentrations are correlated with high concentrations of biogenic VOC precursors.

The most common aldehydes directly released from the tissues of many plants are 2E-hexenal (also called leaf aldehyde) and other C6 aldehydes from the hexenal family (Hatanaka et al., 1987; Arey et al., 1993; Fukui and Doskey, 1998; Kirstine et al., 1998). In undisturbed tissues, hexenal aldehyde emission rates are small (1.0 to 27 ng/g dw h) (Konig et al., 1995). Hexenals function as antibiotics in plant tissues, however, and emission increases in response to physical wounding, herbivory, and pathogen attack, suggesting that current estimates are low.

Formaldehyde and acetaldehyde are directly emitted from plant foliage at relatively low rates (0.2 to 1 μg/g dw h) (Kesselmeier et al., 1997). In most areas, photochemical oxidation of isoprene and other biogenic VOC precursors emitted from vegetation is a more important biogenic source of these compounds than direct emission (e.g., Fried et al., 1997).

Typical background mixing ratios of formaldehyde are 0.1 to 0.15 ppbv (Heikes et al., 1996b). In a rural site in Colorado, the midday background formaldehyde mixing ratio was 1.17 (Fried et al., 1997). In rural areas in Europe, Solberg et al. (1996) observed a seasonal pattern in formaldehyde ambient mixing ratios similar to acetone, where concentrations are highest in the summer (1.3 to 5.9 ppbv), compared to the rest of the year (0.4 to 2.4 ppbv). Arlander et al. (1990) report a latitudinal distribution of formaldehyde from measurements taken over the Pacific Ocean. Maximum mixing ratios (between 0.6 and 0.8 ppbv) of formaldehyde during this cruise were seen between 20°N and the equator, reflecting the latitudinal distribution of both anthropogenic and biogenic alkene precursors.

Organic Acids

Atmospheric mixing ratios of formic and acetic acid typically range from 0.02 to 1.9 in remote and marine locations to 1 to 16 ppbv in urban polluted areas [reviewed in Khare et al. (1999)]. Sources of these organic acids include fossil fuel combustion, biomass burning, direct emissions from formicine ants, soils and plant foliage, and photochemical production in the atmosphere through isoprene and monoterpene oxidation. Direct emissions of both acids have been measured from the European species *Quercus ilex* and *Pinus pinea*, tropical trees in the Amazon, and savanna soils (Kesselmeier et al., 1997; Talbot et al., 1990; Sanhueza and Andreae, 1991). Though the precise biosynthetic mechanisms of organic acids in plant tissues is unknown, acetic acid is formed through lipid metabolism, and formic acid is a by-product of carbohydrate and C1 metabolism [reviewed in Fall (1999)]. In soils, microbial activity is the likely organic acid source.

Vertical profiles and observed seasonal, diurnal, and latitudinal patterns of organic acid concentrations in both precipitation and gas-phase measurements support a significant biogenic source of these acids in rural midlatitude continental, tropical continental, and marine locations (Khare et al., 1999). For example, mixing ratios over the Amazon Basin, and temperate forests in eastern United States were higher during the growing season and the afternoon than in the winter and at night or in the early morning (Keene and Galloway, 1988; Talbot et al., 1988, 1990). Although photochemical production of precursors emitted from vegetation is considered to be the dominant biogenic source of organic acids, direct emission by soils and vegetation can be important in rural areas in the eastern United States and in the Amazon rainforest (Andreae et al., 1988; Talbot et al., 1990, 1995). Formic and acetic acid budgets in marine atmospheres suggest the presence of a natural source as well (Arlander et al., 1990; Heikes et al., 1996a).

3 REGIONAL AND GLOBAL DISTRIBUTION OF BIOGENIC NMVOC EMISSIONS

To understand the impact biogenic NMVOCs have on tropospheric chemistry, reliable emission estimates at local, regional, and global scales are necessary. Flux measurements made at a variety of scales are the primary means of both developing and evaluating these emission estimates. The techniques used to measure these fluxes are reviewed in Guenther et al. (1996). Enclosure methods are used to estimate fluxes on small scales including from a single leaf, branch, or whole tree. These measurements are a particularly good way to quantify species-specific basal emission rates (or the capacity to emit NMVOCs under a standard set of environmental conditions). Tower-based micrometeorological techniques are used to directly measure canopy-scale fluxes of NMVOCs on diurnal, seasonal, and annual time scales. These techniques include eddy covariance, relaxed eddy accumulation (REA), surface layer gradient, and tracer methods. Finally, sampling systems on tethered balloons and aircraft are used to construct vertical mixing ratio profiles and calculate surface fluxes on scales of tens to hundreds of kilometers using

eddy accumulation, REA, mixed-layer mass balance, and mixed-layer gradient methods.

To construct inventories, basal emission rates from a wide range of vegetation classes are modified by instantaneous changes in both temperature and light intensity using algorithms developed by Guenther et al. (1993), multiplied by estimates of foliar density of each vegetation class, and aggregated to give flux estimates on regional and global scales (Lamb et al., 1987; Guenther et al., 1995; Guenther, 1997). Large uncertainties are associated with these inventories, however, due to gaps in our knowledge of (1) the contribution of nonfoliar emissions, (2) physiological and ecological controls over emissions from plants, (3) specific emission factors from a wider variety of plants and ecosystems, particularly of nonterpenoid NMVOCs, and (4) detailed data on coverage of ecosystem type, foliage density, surface temperatures, and radiation properties (Steinbrecher, 1997).

These inventories are useful for identifying where, on a regional basis, biogenic contributions to total NMVOC fluxes are particularly relevant due to vegetation type, foliar density, and ambient temperature patterns. For example, in urban areas such as Los Angeles, biogenic NMVOCs contribute a relatively small fraction to the total VOC emissions (Benjamin et al., 1997). In Atlanta and remote rural areas, however, biogenic sources, during the summer months especially, can dominate the total VOC emission profile (Geron et al., 1995; Hagerman et al., 1997). In North America as a whole, and in Norway, Sweden, and Finland, biogenic emissions of VOCs exceed anthropogenic emissions (Guenther et al., 1995; Simpson et al., 1995). In Italy, biogenic emissions account for 50% of the total VOCs emitted (Simpson et al., 1995).

On a global scale, Guenther et al. (1995) derived estimates for emissions of isoprene (420 Tg C/yr), monoterpenes (130 Tg C/yr), and other reactive VOCs (280 Tg C/yr). As expected due to the influence of light intensity and temperature on emissions, biogenic fluxes show seasonal as well as latitudinal differences (Guenther et al., 1995; Guenther, 1999). Drought deciduous forests and savannas in the tropics contributed half of all the global VOCs from biogenic sources in this estimate. Other woodlands, crops, and shrublands contributed 10 to 20% of these fluxes. Crops, in particular, were high emitters of VOCs other than isoprene and monoterpenes.

Relative to isoprene, trends in the global distribution of monoterpene and other NMVOC fluxes are hard to find. The high reactivity, spatial and temporal variability in source strengths, and uncertainties in reliable identification and quantification of these species have contributed to large variability in observed ambient mixing ratios. Thus, although measurements of these species have been made, considerable work is necessary to truly understand the global distribution of these reactive VOCs in the atmosphere.

4 SUMMARY AND CONCLUSIONS

A variety of nonmethane hydrocarbons are released from natural sources, particularly plant foliage, in quantities sufficient to alter production of tropospheric ozone,

organic acids and nitrates, PAN, OH, and CO. Isoprene and monoterpenes together dominate NMVOC fluxes from many species, ecosystems, regions, and on a global scale. Although the mechanisms of production, emission, and degradation in the atmosphere are fairly well known for isoprene and monoterpenes, uncertainties such as why only certain plants produce these compounds and detailed distributions of the vegetation sources remain.

Uncertainties are largest for VOCs other than isoprene and monoterpenes. Nonterpenoid hydrocarbons contribute an estimated 45% of the total biogenic VOC global fluxes. To make more reliable estimates of the source strength and atmospheric impacts of these hydrocarbons, a better understanding of the biological and ecological factors that control spatial and temporal variability in these fluxes is needed.

Current NMVOC inventories rely on empirical models based only on the response of emissions to temperature, light, and foliar density. Given that emissions of isoprene, monoterpenes, and many of the other VOCs are influenced by a whole suite of physiological and ecological factors, using these inventories to predict emissions in response to disturbances, land-use change, and/or climate change is risky. An important aspect of future biogenic VOC research is to incorporate a more mechanistic understanding of VOC production and emission into emission inventories (Monson et al., 1995). In this way inventories will be able to extrapolate emission rates and the impacts of these emissions on atmospheric chemistry across complex ecological gradients in both space and time.

REFERENCES

Abeles, F. B., P. W. Morgan, and M. E. Saltveit, *Ethylene in Plant Biology*, 2nd ed., Academic, New York, 1992.

Adam, K.-P., J. Crock, and R. Croteau, Partial purification and characterization of a monoterpene cyclase, limonene synthase, from the liverwort *Ricciocarpos natans*, *Arch. Biochem. Biophys. 332*, 352–356, 1996.

Anderson, L. J., P. C. Harley, R. K. Monson, and R. B. Jackson, Reduction of isoprene emissions from live oak (Quercus fusiformis) with oak wilt, *Tree Phys., 20*, 1199–1203, 2000.

Andreae, M. O., R. W. Talbot, T. W. Andreae, and R. C. Harriss, Formic and acetic acid over the Central Amazon region, Brazil. 1. Dry season, *J. Geophys. Res., 93*, 1616–1624, 1988.

Andreae, M. O., and P. J. Crutzen, Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry, *Science, 276*, 1052–1058, 1997.

Arey, J., A. M. Winer, R. Atkinson, S. M. Aschmann, W. D. Long, and C. L. Morrison, The emission of (*Z*)-3-hexen-1-ol, (*Z*)-3-hexenylaceteate and other oxygenated hydrocarbons from agricultural plant species, *Atmos. Environ., 25A*, 1063–1075, 1993.

Arlander, D. W., D. R. Cronn, J. C. Farmer, F. A. Menzi, and H. H. Westberg, Gaseous oxygenated hydrocarbons in the remote marine troposphere, *J. Geophys. Res., 95*, 16391–16403, 1990.

Banthorpe, D., and V. Charlwood, The terpenoids, in E. Bell, and V. Charlwood (Eds.), *Encyclopedia of Plant Physiology*, Springer-Verlag, Berlin, 1980, pp. 185–220.

Benjamin, M. T., M. Sudol, D. Vorsatz, and A. M. Winer, A spatially and temporally resolved biogenic hydrocarbon emissions inventory for the California South coast air basin, *Atmos. Environ.*, *31*, 3087–3100, 1997.

Bertin, N., and M. Staudt, Effect of water stress on monoterpene emissions from young potted Holm oak (*Quercus ilex* L.) trees, *Oecologia*, *107*, 456–462, 1996.

Bonsang, B., and C. Boissard, Global distribution of reactive hydrocarbons in the atmosphere, in C. N. Hewitt (Ed.), *Reactive Hydrocarbons in the Atmosphere*, Academic, San Diego, 1999, pp. 43–97.

Cao, X. L., C. Boissard, A. J. Juan, C. N. Hewitt, and M. Gallagher, Biogenic emissions of volatile organic compounds from gorse (*Ulex europaeus*): Diurnal emission fluxes at Kelling Heath, England, *J. Geophys. Res.*, *102*, 18903–18915, 1997.

Chameides, W. L., R. W. Lindsay, J. Richardson, and C. S. Kiang, The role of biogenic hydrocarbons in urban photochemical smog: Atlanta as a case study, *Science*, *241*, 1–10, 1988.

Cleveland, C. C., and J. B. Yavitt, Microbial consumption of atmopheric isoprene in a temperate forest soil, *Appl. Environ. Microbiol.*, *64*, 172–177, 1998.

Conrad, R., Soil microbial processes and the cycling of atmospheric trace gases, *Phil. Trans. R. Soc. Lond. Ser. A.*, *351*, 219–230, 1995.

Dement, W. A., B. J. Tyson, and H. A. Mooney, Mechanism of monoterpene volatilization in *Salvia mellifera*, *Phytochemistry*, *14*, 2555–2557, 1975.

Fall, R., Biogenic emissions of volatile organic compounds from higher plants, in C. N. Hewitt (Ed.), *Reactive Hydrocarbons in the Atmosphere*, Academic, San Diego, 1999, pp. 43–97.

Fehsenfeld, F., J. Calvert, R. Fall, P. Goldan, A. Guenther, C. N. Hewitt, B. Lamb, S. Liu, M. Trainer, H. Westberg, and P. Zimmerman, Emissions of volatile organic compounds from vegetation and the implications for atmospheric chemistry, *Global Biogeochem. Cycles*, *6*, 389–430, 1992.

Fried, A., S. Mckeen, S. Sewell, J. Harder, B. Henry, P. Goldan, W. Kuster, E. Williams, K. Baumann, R. Shetter, and C. Cantrell, Photochemistry of formaldehyde during the 1993 Tropospheric OH Photochemistry Experiment, *J. Geophys. Res.*, *102*, 6283–6296, 1997.

Fukui, Y., and P. V. Doskey, Air-surface exchange of nonmethane organic compounds at a grassland site: Seasonal variations and stressed emissions, *J. Geophys. Res.*, *103*, 13153–13168, 1998.

Geron, C., T. Pierce, and A. Guenther, Reassessment of biogenic volatile organic compound emissions in the Atlanta area, *Atmos. Environ.*, *29*, 1569–1578, 1995.

Gershenzon, J., and R. Croteau, Terpenoids, in G. A. Rosenthal and M. R. Gerenbaum (Eds.), *Herbivores. Their Interactions with Secondary Plant Metabolites*, 2nd ed., Vol. 1: *The Chemical Participants*, Academic, San Diego, 1991, pp. 165–219.

Goldan, P. D., W. C. Kuster, and F. C. Fehsenfeld, Nonmethane hydrocarbon measurements during the Tropospheric OH Photochemistry Experiment, *J. Geophys. Res.*, *102*, 6315–6324, 1997.

Goldan, P. D., W. C. Kuster, F. C. Fehsenfeld, and S. A. Montzka, The observation of a C5 alcohol in a North American pine forest, *Geophys. Res. Lett.*, *20*, 1039–1042, 1993.

Goldan, P. D., W. C. Kuster, F. C. Fehsenfeld, and S. A. Montzka, Hydrocarbon measurements in the southeastern United States: The Rural Oxidants in the Southern Environment (ROSE) Program 1990, 1995.

Goldstein, A. H., S. M. Fan, M. L. Goulden, J. W. Munger, and S. C. Wofsy, Emissions of ethene, propene, and 1-butene by a midlatitude forest, *J. Geophys. Res., 101*, 9149–9157, 1996.

Goldstein, A. H., and G. W. Schade, Quantifying the biogenic and anthropogenic contributions to high concentrations of acetone observed in the Sierra Nevada Mountains (CA), paper presented at American Geophysical Union Fall Meeting, San Francisco, CA, December 13–17, 1999.

Graedel, T. E., *Chemical Compounds in the Atmosphere*, Academic, New York, 1979.

Guenther, A., Seasonal and spatial variations in natural volatile organic compound emissions, *Ecol. Appl., 7*, 34–45, 1997.

Guenther, A., Modeling biogenic VOC emissions to the atmosphere, in C. N. Hewitt (Ed.), *Reactive Hydrocarbons in the Atmosphere*, Academic, San Diego, 1999, pp. 97–118.

Guenther, A., P. Zimmerman, P. Harley, R. Monson, and R. Fall, Isoprene and monoterpene emission rate variability: Model evaluation and sensitivity analysis, *J. Geophys. Res., 98*, 12609–12617, 1993.

Guenther, A., P. Zimmerman, and M. Wildermuth, Natural volatile organic compound emission rate estimates for U.S. woodland landscapes, *Atmos. Environ., 28*, 1197–1210, 1994.

Guenther, A., C. N. Hewitt, D. Erickson, R. Fall, C. Geron, T. Gradel, P. Harley, L. Klinger, M. Lerdau, W. A. McKay, T. Pierce, B. Scholes, R. Steinbrecher, R. Tallamraju, J. Taylor, and P. Zimmerman, A global model of natural volatile organic compound emissions, *J. Geophys. Res., 100*, 8873–8892, 1995.

Guenther, A., W. Baugh, K. David, G. Hamptom, P. Harley, L. Klinger, P. Zimmerman, E. Allwine, S. Dilts, B. Lamb, H. Westberg, D. Baldocchi, C. Geron, and T. Pierce, Isoprene fluxes measured by enclosure, relaxed eddy accumulation, surface-layer gradient, mixed-layer gradient, and mass balance techniques, *J. Geophys. Res., 101*, 18555–18568, 1996.

Haagen-Smit, A. J., Chemistry and physiology of Los Angeles smog, *Ind. Eng. Chem., 44*, 1342–1345, 1952.

Hagerman, L. M., V. P. Aneja, and W. A. Lonneman, Characterization of nonmethane hydrocarbons in the rural Southeast United States, *Atmos. Environ., 31*, 4017–4038, 1997.

Harley, P. C., M. E. Litvak, T. D. Sharkey, and R. K. Monson, Isoprene emissions from velvelt bean leaves – interactions among nitrogen availability, growth photon flux density and leaf development, *Plant Phys., 105*, 279–285, 1994.

Harley, P., V. Fridd-Stroud, J. Greenberg, A. Guenther, and P. Vasconcellos, Emission of 2-methyl-3-buten-2-ol by pines: A potentially large natural source of reactive carbon to the atmosphere, *J. Geophys. Res., 103*, 25479–25486, 1998.

Harley, P., R. K. Monson, and M. T. Lerdau, Ecological and evolutionary aspects of isoprene emission from plants, *Oecologia, 118*, 109–123, 1999.

Hatanaka, A., T. Kajiwara, and J. Sekiya, Biosynthetic pathways for C6-aldehydes formation from linolenic acid in greed leaves, *Chem. Phys. Lipids, 44*, 341–361, 1987.

Heikes, B., M. Lee, D. Jacob, R. Talbot, J. Bradshaw, H. Singh, D. Blake, B. Anderson, H. Fuelberg, and A. M. Thompson, Ozone, hydroperoxides, oxides of nitrogen, and hydrocarbon budgets in the marine boundary layer over the South Atlantic, *J. Geophys. Res., 101*, 24221–24234, 1996a.

Heikes, B., B. McCuly, X. Zhou, Y.-N. Lee, K. I. Mopper, X. Chen, G. Mackay, D. Karecki, H. Schiff, T. Campos, and E. Atlas, Formaldehyde methods comparison in the remote lower troposphere during the Mauna Loa Photochemistry Experiment 2, *101*, 14741–14755, 1996b.

Helmig, D., L. F. Klinger, A. Guenthe, L. Vierling, C. Geron and P. Zimmerman, Biogenic volatile organic compound emissions (BVOCs) I. Identifications from three continental sites in the US, *Chemosphere, 38*, 2163–2187, 1999.

Helmig, D., B. Balsley, K. Davis, L. R. Kuck, M. Jensen, J. Bognar, T. Smith, Jr., R. Vasquez Arrieta, R. Rodriguez, and J. W. Birks, Vertical profiling and determination of landscape fluxes of biogenic nonmethane hydrocarbons within the planetary boundary layer in the Peruvian Amazon, *J. Geophys. Res.*, *103*, 25519–25532, 1998.

Isidorov, V. A., I. G. Zenkevich, and B. V. Ioffe, Volatile organic compounds in the atmosphere of forests, *Atmos. Environ.*, *19*, 1–8, 1985.

Jacob, D. J., and S. C. Wofsy, Photochemistry of biogenic emissions over the Amazon forest, *J. Geophys. Res.*, *93*(D2), 1477–1486, 1988.

Karl, T., A. Guenther, C. Lindinger, A. Jordan, R. Fall, and W. Lindinger. Eddy covariance measurements of oxygenated volatile organic compound fluxes from crop harvesting using a redesigned proton-transfer-reaction mass spectrometer. *J. Geophys. Res.*, *106*, 24157–24167, 2001.

Kasting, J. F., and H. B. Singh, Nonmethane hydrocarbons in the troposphere – impact on the odd hydrogen and odd nitrogen chemistry, *J. Geophys. Res., 91*, 3239–3256, 1986.

Keene, W. C., and J. N. Galloway, The biogeochemical cycling of formic and acetic acids through the troposphere: An overview of current understanding, *Tellus*, *40B*, 322–334, 1988.

Kesselmeier, J., K. Bode, U. Hofmann, H. Muller, L. Schaefer, A. Wolf, P. Ciccioli, E. Brancaleoni, A. Cecinato, M. Frattoni, P. Foster, C. Ferrari, V. Jacob, J. L. Fugit, L. Dutaur, V. Simon, and L. Torres, Emission of short chained organic acids, aldehydes and monoterpenes from *Quercus ilex* L. and *Pinus pinea* L. in relation to physiological activities, carbon budget and emission algorithms, *Atmos. Environ.*, *31*(SI), 119–133, 1997.

Khare, P., N. Kumar, K. M. Kumari, and S. S. Srivastava, Atmospheric formic and acetic acids: An overview, *Rev. Geophys.*, *37*, 227–248, 1999.

Kirstine, W., I. Galbally, Y. Yuerong, and M. Hooper, Emissions of volatile organic compounds (primarily oxygenated species) from pasture, *J. Geophys. Res.*, *103*, 10605–10619, 1998.

König, G., M. Brunda, H. Puxbaum, C. N. Hewitt, and S. C. Duckham, Relative contribution of oxygenated hydrocarbons to the total biogenic VOC emissions of selected mid-European agricultural and natural plant species, *Atmos. Environ.*, *29*, 861–874, 1995.

Kotzias, D., C. Konidari, and C. Sparta, Volatile carbonyl compounds of biogenic origin—Emission and concentration in the atmosphere, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Carbon compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 67–78.

Lamanna, M. S., and A. H. Goldstein, In situ measurements of C2–C10 volatile organic compounds above a Sierra Nevada ponderosa pine plantation, *J. Geophys. Res.*, *104*, 21247–21262, 1999.

Lamb, B., A. Guenther, D. Gay, and H. Westberg, A national inventory of biogenic hydrocarbon emissions, *Atmos. Environ.*, *21*, 1695–1705, 1987.

Lamb, B., H. Westberg, and G. Allwine, Biogenic hydrocarbon emissions from deciduous and coniferous trees in the United States, *J. Geophys. Res.*, *90*, 2380–2390, 1985.

Langenheim, J. H., Higher plant terpenoids: A phytocentric overview of their ecological roles, *J. Chem. Ecol.*, *20*, 1223–1280, 1994.

Lerdau, M., and H. L. Throop, Sources of variability in isoprene emission and photosynthesis in two species of tropical wet forest trees, *Biotropica*, *32*, 670–676, 2000.

Lerdau, M., M. Litvak, P. Palmer, and R. Monson, Controls over monoterpene emissions from boreal forest conifers, *Tree Phys.*, *17*, 491–499, 1997.

Lerdau, M., P. Matson, R. Fall, and R. Monson, Ecological controls over monoterpene emission from Douglas-fir *Pseudotusga menziesii*, *Ecology*, *76*, 2640–2647, 1995.

Lichtenthaler, H. K., J. Schwender, A. Disch, and M. Rohmer, Biosynthesis of isoprenoids in higher plant chloroplasts proceeds via a mevalonate-independent pathway, *FEBS Lett.*, *400*, 271–274, 1997.

Lindskog, A., The influence of biosphere on the budgets of VOC: Ethane, propane, *n*-butane and *i*-butane, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Carbon Compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 45–52.

Litvak, M. L., S. Madronich, and R. K. Monson, Herbivore-induced monoterpene emissions from coniferous forests: Potential impact on local tropospheric chemistry, *Ecol. Appl.*, *9*, 1147–1159, 1999.

Litvak, M. E., and R. K. Monson, Patterns of constitutive and induced monoterpene production in conifer needles in relation to insect herbivory, *Oecologia*, *118*, 531–540, 1998.

Loreto, F., P. Ciccioli, A. Cecinato, E. Brancaleoni, M. Frattoni, C. Fabozzi, and D. Tricoli, Evidence of the photosynthetic origin of monoterpenes emitted by *Quercus ilex* L. leaves by C^{13} labeling, *Plant Physiol.*, *110*, 1317–1322, 1996.

MacDonald, R. C., and R. Fall, Detection of substantial emissions of methanol from plants to the atmosphere, *Atmos. Environ.*, *27A*, 1709–1713, 1993.

MacDonald, R. C., T. W. Kimmerer, and M. Razzaghi, Aerobic ethanol production by leaves: Evidence for air pollution stress in trees in the Ohio River Valley, USA, *Environ. Pollut.*, *62*, 337–351, 1989.

Monson, R. K., P. C. Harley, M. E. Litvak, M. Wildermuth, A. B. Guenther, P. R. Zimmerman, and R. Fall, Environmental and developmental controls over the seasonal pattern of isoprene emission from aspen leaves. *Oecologia, 99*, 260–270, 1994.

Monson, R. K., M. T. Lerdau, T. D. Sharkey, D. S. Schimel, and R. Fall, Biological aspects of constructing volatile organic compound emission inventories, *Atmos. Env.*, *29*, 2989–3002, 1995.

Müller, J. F., Geographical distribution and seasonal variation of surface emissions and deposition velocities of atmospheric trace gases, *J. Geophys. Res.*, *97*, 3787–3804, 1992.

Nemecek-Marhsall, M., R. C. MacDonald, J. F. Franzen, C. L. Wojciechowski, and R. Fall, Methanol emission from leaves, *Plant Phys.*, *108*, 1359–1368, 1995.

Pandis, S. N., S. E. Paulson, J. H. Seinfeld, and R. C. Flagan, Aerosol formation in the photooxidation of isoprene and β-pinene, *Atmos. Environ.*, *Part A*, *26*, 2269–2282, 1991.

Puxbaum, H., Biogenic emissions of alcohols, ester, ether and higher aldehydes, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 79–99.

Rasmussen, R. A., and M. A. Khalil, Isoprene over the Amazon basin, *J. Geophys. Res.*, *93*, 1417–1421, 1988.

Ratte, M., C. Plassdulmer, R. Koppmann, and J. Rudolph, Horizontal and vertical profiles of light hydrocarbons in sea water related to biological, chemical and physical parameters, *Tellus, Series B*, *47*, 607–623, 1995.

Roberts, J. M., F. C. Fehsenfeld, D. L. Albritton, and R. E. Sievers, Measurement of monoterpene hydrocarbons at Niwot Ridge, Colorado, *J. Geophys. Res.*, *88*, 10667–10678, 1983.

Roselle, S. J., T. E. Pierce, and K. L. Schere, The sensitivity of regional ozone modeling to biogenic hydrocarbons, *J. Geophys. Res.*, *96*, 7371–7394, 1991.

Rudolph, J., Biogenic sources of atmospheric alkenes and acetylene, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Carbon Compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 53–65.

Sanadze, G. A., Isoprene effect–light dependent emission of isoprene by green parts of plants, in T. D. Sharkey, E. A. Holland, and H. A. Mooney (Eds.), *Trace Gas emissions by Plants*, Academic, San Diego, 1991, pp. 135–152.

Sanhueza, E., and M. O. Andreae, Emission of formic and acetic acids from tropical savanna soils, *Geophys. Res. Lett.*, *18*, 1707–1710, 1991.

Sawada, S., and T. Totsuka, Natural and anthropogenic sources and fate of atmospheric ethylene, *Atmos. Environ.*, *20*, 821–832, 1986.

Sharkey, T. D., Emission of low molecular mass hydrocarbons from plants, *Trends Plant Sci.*, *1*, 78–82, 1996.

Sharkey, T. D., Isoprene production in trees, in H. Rennenberg, W. Eschrich, and H. Ziegler (Eds.), *Trees—Contributions to Modern Tree Physiology*, Backhuys, The Netherlands, 1997, pp. 109–118.

Sharkey, T. D., E. L. Singsaas, M. T. Lerdau, and C. D. Geron, Weather effects on isoprene emission capacity and applications in emissions algorithms, *Ecol. Appl.*, *9*, 1132–1137, 1999.

Sillman, S., and F. J. Samson, Impact of temperature on oxidant photochemistry in urban, polluted rural and remote environments, *J. Geophys. Res.*, *100*, 11497–11508, 1995.

Simpson, D., A. Guenther, C. N. Hewitt, and R. Steinbrecher, Biogenic emissions in Europe. I. Estimates and uncertainties, *J. Geophys. Res.*, *100*, 506–512, 1995.

Singh, H. B., M. Kanakidou, P. J. Crutzen, and D. J. Jacob, High concentrations and photochemical fate of oxygenated hydrocarbons in the global troposphere, *Nature*, *378*, 50–54, 1995.

Singh, H. B., and P. R. Zimmerman, Atmopheric distribution and sources of nonmethane hydrocarbons, in J. O. Nriagu (Ed.), *Gaseous Pollutants: Characterization and Cycling*, Wiley-Interscience, New York, 1992, pp. 177–235.

Solberg, S., C. Dye, N. Schmidbauer, A. Herzog, and R. Gehrig, Carbonyls and nonmethane hydrocarbons at rural European sites from the Mediterranean to the arctic, *J. Atmos. Chem.*, *25*, 33–66, 1996.

Staudt, M., N. Bertin, U. Hansen, G. Seufert, P. Ciccioli, P. Foster, B. Frenzel, J.-L. Fugit, and L. Torres, The BEMA-project: Seasonal and diurnal patterns of monoterpene emissions from *Pinus pinea* (L.) measured under field conditions, *Atmos. Environ.*, *31*, 145–156, 1997.

Steinbrecher, R., Isoprene: Production by plants and ecosystem-level estimates, in G. Helas, J. Slanina, and R. Steinbrecher (Eds.), *Biogenic Volatile Organic Carbon Compounds in the Atmosphere*, SPB Academic, Amsterdam, 1997, pp. 101–114.

Steinbrecher, R., W. Schürmann, A.-M. Schreiner, and H. Ziegler, Terpenoid emissions from common oak (*Quercus robur* L.) and Norway spruce (*Picea abies* L. Karst.), in J. Slanina, G. Angeletti, and S. Beilke (Eds.), *Proceedings of the Joint CEC/BIATEX Workshop on the General Assessment of Biogenic Emissions and Deposition of Nitrogen Compounds, Sulfur Compounds and Oxidants in Europe*, CEC Environ. Res. Progr. Report 47, 1993, pp. 251–261.

Talbot, R. W., B. W. Mosher, B. G. Heikes, D. J. Jacob, J. W. Munger, B. C. Daube, W. C. Keene, J. R. Maben, and R. S. Artz, Carboxylic-acids in the rural continental atmosphere over the eastern United States during the Shenandoah cloud and photochemistry experiment, *J. Geophys. Res., 100*, 9335–9343, 1995.

Talbot, R. W., M. O. Andreae, H. Berresheim, D. J. Jacob, and K. M. Beecher, Sources and sinks of formic, acetic, and pyruvic acids over central Amazonia, 2, Wet season, *J. Geophys. Res., 95*, 16799–16811, 1990.

Talbot, R. W., K. M. Beecher, R. C. Harriss, and W. R. Cofer III, Atmospheric geochemistry of formic and acetic acids at a midlatitude temperate site, *J. Geophys. Res., 93*, 1638–1652, 1988.

Talbot, R. W., B. W. Mosher, B. G. Heikes, D. J. Jacob, J. W. Munger, B. C. Daube, W. C. Keene, J. R. Maben, and R. S. Artz, Carboxylic acids in the rural continental atmosphere over the eastern United States during the Shenandoah Cloud and Photochemistry Experiment.

Tingey, D. T., D. P. Turner, and J. A. Weber, Factors controlling the emissions of monoterpenes and other volatile organic compounds, in T. D. Sharkey, E. A. Holland, and H. Mooney (Eds.), *Trace Gas Emissions from Plants*, Academic, San Diego, 1991, pp. 93–119.

Trainer, M., E. J. Williams, D. D. Parrish, M. P. Buhr, E. J. Allwine, H. Westberg, F. C. Fehsenfeld, and S. C. Liu, Models and observations of the impact of natural hydrocarbons on rural ozone.

Yani, A., G. Pauly, M. Faye, F. Salin, and M. Gleizes, The effect of a long term water stress on the metabolism and emission of terpenes of the foliage of *Cupressus sempervirens*, *Plant Cell Environ., 16*, 975–981, 1993.

Warneke, C., T. Karl, H. Judmaier, A. Hansel, A. Jordan, W. Lindinger, and P. Crutzen, Acetone, methanol and other partially oxidized volatile organic emissions from dead plant matter by abiological processes: Significance for atmospheric HO_x chemistry, *Global Biogeochem. Cycles, 13*, 9–17, 1999.

Winer, A., J. Arey, R. Atkinson, S. Aschman, W. Long, L. Morrison, and D. Olszyk, Emission rates of organics from vegetation in California's Central Valley, *Atmos. Environ., 26A*, 2647–2659, 1992.

CHAPTER 8

ATMOSPHERIC SULFUR

D. D. DAVIS, G. CHEN, AND M. CHIN

1 INTRODUCTION

The focus of this chapter is that of providing the reader with an overview of atmospheric sulfur. It will address the issues of where sulfur comes from, how it is processed, and how it gets returned to the planetary surface. It will also endeavor to show how sulfur, during its atmospheric cycle, plays a significant role in helping to maintain a stable global environment.

Sulfur is an element that is essential to life on this planet. Living organisms at nearly all levels of sophistication ingest sulfur from their environment, mainly in the form of sulfate or amino acid sulfur. But living organisms not only ingest sulfur, they also have a decisive impact on the chemical forms and total burden that is found in the atmosphere. During the process known as *assimilatory sulfate reduction*, microorganisms and plants use sulfate to build sulfur-containing proteins for purposes of storing energy and to support cell growth. During food digestion, animals are able to generate energy from the catabolism of these proteins, breaking them down to their chemical building blocks, the amino acids. The further breakdown of these compounds results in the release of volatile sulfur back to the environment.

Some microorganisms living in anoxic environments, such as tidal flats, obtain energy from using sulfate as an electron acceptor instead of O_2. This process is called *disimilatory sulfate reduction*. Hydrogen sulfide (H_2S) released during this process often combines with iron minerals to form pyrite, FeS, resulting in its incorporation into sediment layers. Alternatively, the H_2S may react with buried organic matter, thus forming a source of sulfur in fossil oil and coal deposits. In general, the turnover of sulfur in dissimilatory processes is several orders of magnitude faster than in assimilatory processes. The biological sulfur cycle is therefore mainly controlled by anaerobic sulfate reducing bacteria. [For further details on the

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

assimilatory and dissimilatory biological processes, the reader is referred to reviews by Krouse and McCready (1979) and Andreae and Jaeschke (1992).]

Sulfur, having six valence electrons, has the potential for existing in the atmosphere in a wide range of oxidation states, ranging from -2 to $+6$, with the most common states being -2, $+4$, and $+6$. In most remote global locations, atmospheric sulfur is found at concentration levels of only 1 ppbv (part-per billion by volume) or less. However, for continental regions, particularly those experiencing significant industrial development, concentrations can reach upwards of 100 to 200 ppbv. Similar levels can be found in regions under the influence of active volcanoes. This trend in global sulfur levels is a strong reflection of the distribution of the major sources of sulfur as illustrated in Figure 1. From this abbreviated picture of the atmospheric sulfur cycle, the most critical members of the atmospheric sulfur family are identified as dimethyl sulfide (DMS), sulfur dioxide (SO_2), sulfuric acid (H_2SO_4), and aerosol sulfate (SO_4^{2-}). The latter species is typically found in the form of condensation nuclei (CN) or cloud condensation nuclei (CCN). Two of the major primary sulfur sources are shown as SO_2, emitted from the burning of large quantities of fossil fuel or, alternatively, from volcanoes, and DMS, which predominantly is released from the world's oceans. This source is obviously dispersed over a much larger global surface area than is primary SO_2, leading to much lower concentration levels of sulfur over remote regions. In the latter context, shown also in Figure 1, is the most recently identified remote sulfur source, emissions from ships (Corbett and Fischbeck, 1997; Corbett et al., 1999). This new source, however, is significantly smaller than the three previously discussed.

Among the central points revealed in Figure 1 is the fact that the atmosphere can be viewed as a large oxidizing chemical reactor in which sulfur, emitted from Earth's surface, enters the atmosphere in a chemically reduced oxidation state (typically -2 and $+4$) is oxidized to the $+6$ state, and then in ionic form (i.e., higher solubility) is returned to the biosphere, thus closing the cycle. The processes responsible for oxidizing sulfur are shown as occurring by both gas phase as well as heterogeneous reactions. Once in the $+6$ oxidation state, this sets the stage for the final contribution from atmospheric sulfur toward maintaining a stable global environment, namely, its impact on the planetary radiation budget. As shown in Figure 1 atmospheric aerosols, which are predominately composed of sulfur, can have a significant impact on the planet's climate via their influence on direct scattering of incoming solar radiation and by their controlling the radiative characteristics and formation rates of clouds (Charlson et al., 1992).

In Section 2 of this chapter, we will expand on the source inventories for DMS and SO_2 as well as present inventories for several less important primary sulfur source species, including those for H_2S, carbony sulfide (OCS), and carbon disulfide (CS_2). Of particular significance will be the hemispheric distributions of these collective sulfur sources and how they manifest themselves in observed concentration levels. Section 3 will also explore in greater detail the oxidation processes responsible for converting the dominant sulfur species (i.e., DMS and SO_2) into forms that result in their removal. Section 4 will combine the source inventory data presented in Section 3 and the chemical transformation information discussed in Section 2 in exploring global distributions of SO_2 and SO_4^{2-}. Finally, in Section 5,

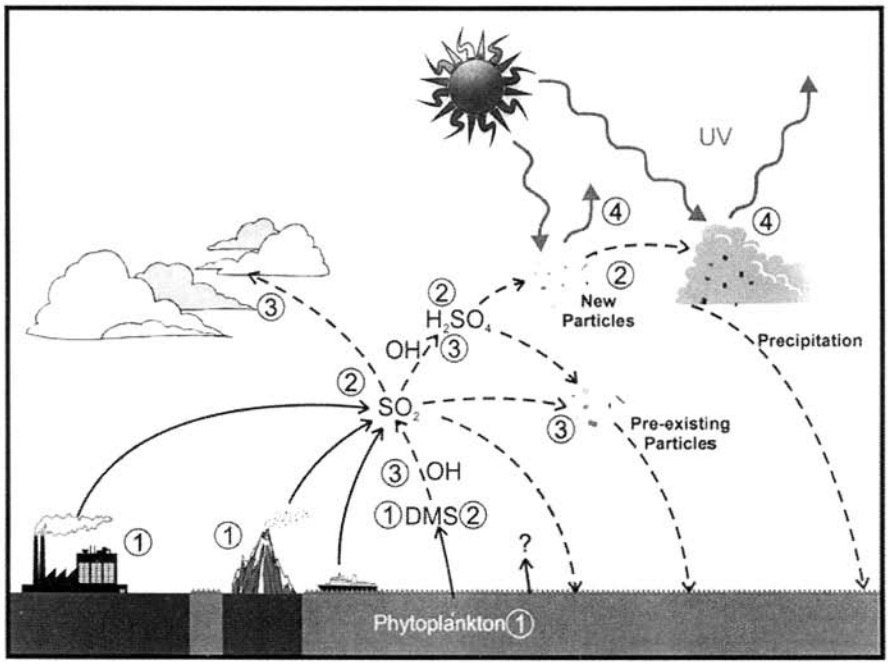


Figure 1 Simplified tropospheric sulfur cycle: ① The three largest documented global sulfur sources: ocean emissions, volcanoes, and fossil fuel burning. ② The most critical species involved in the cycling of sulfur: DMS, SO_2, H_2SO_4, and SO_4^{2-} (as CN, and CCN). ③ The major chemical processes in the cycling of tropospheric sulfur encompass gas-phase and heterogeneous reactions. ④ Among the important environmental impacts of tropospheric sulfur: formation of new particles and promotion of aerosol growth. Both are critical factors in Earth's radiation budget.

we present an overview of sources, sinks, and transformations of sulfur in the stratosphere with a special emphasis on sulfur sources responsible for maintaining the "background" level of stratospheric aerosol.

The authors note that because of the more fundamental chemical nature of Section 3, the discussion in this section is necessarily presented in greater detail than are other sections. The reader may choose, therefore, to by-pass this section. As the text is configured, this can be done without a major loss in grasping the larger global picture of atmospheric sulfur and how this element is critically coupled to the larger planetary environment.

2 CHEMICAL FORMS, SOURCES, AND CONCENTRATION LEVELS

As shown in Figure 2, some 11 different sulfur compounds define over 98% of the sulfur speciation in the atmosphere. Those in which sulfur is found bonded to either

Marine Sulfur Species

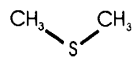
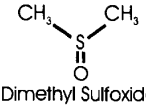
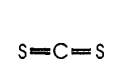
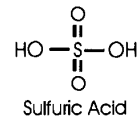
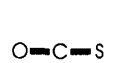
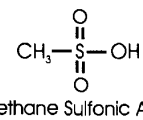
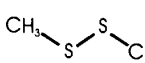
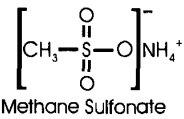
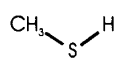
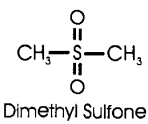
Structure and Name	Symbol	Structure and Name	Symbol
 Hydrogen Sulfide	(H_2S)	 Sulfur Dioxide	(SO_2)
 Dimethylsulfide	(DMS)	 Dimethyl Sulfoxide	(DMSO)
 Carbon Disulfide	(CS_2)	 Sulfuric Acid	(H_2SO_4)
 Carbonyl Sulfide	(OCS)	 Methane Sulfonic Acid	(MSA)
 Dimethyl Disulfide	(DMDS)	 Methane Sulfonate	(MS)
 Methyl Mercaptan	(MeSH)	 Dimethyl Sulfone	($DMSO_2$)

Figure 2 Chemical formulas and simple structures of the most common sulfur species in the troposphere. This list defines ~98% of the sulfur loading of the atmosphere.

hydrogen or carbon are typically in the lowest oxidation state, e.g., −2. As oxygen is sequentially added to sulfur, the oxidation state moves to 0, +4, and +6. Examples of these different states include dimethyl sulfoxide (DMSO), SO_2, and methane sulfonic acid (CH_3SO_3H), respectively.

As noted in Section 1, of those sulfur species shown in Figure 2, SO_2 and DMS are by far the most important primary forms emitted into the atmosphere. This point is further illustrated in Table 1, which provides a compilation of primary sulfur sources. From here it can be seen that of the total global average flux of 128 Tg S/yr, nearly 90% of this is defined by SO_2 and DMS emissions.

TABLE 1 Global Sulfur Emission Inventory

	DMS	SO_2	SO_4^{2-}	Other Reduced Sulfur	Total
Northern Hemisphere					
Combustion of fossil fuel	0.37–0.42	65–90	1.8–2	1.1–1.4	68–94
Oceans	5.8–9.7			0.1–0.4	5.9–10
Volcanoes		2.4–6.6	1.4–2.9	0.47–1.2	4.3–11
Other[a]	0.032–0.6	1.3–1.6	1.3–2.5	0.18–1.6	2.8–6.2
Anthropogenic total	0.37–0.42	65–90	1.9–2.1	1.1–1.5	70–96
Natural total	5.8–10	2.4–6.6	2.6–5.3	0.7–3	12–25
N.H. total	6.2–11	69–98	4.5–7.4	1.8–4.5	81–121
Southern Hemisphere					
Combustion of fossil fuel	0.02	7.1–9.2	0.2	0.12–0.13	7.4–9.6
Oceans	9.2–15			0.04–0.36	9.2–15
Volcanoes		1–2.6	0.6–1.1	0.13–0.43	1.7–4.1
Other[a]	0.021–0.2	1–1.3	0.8–1.6	0.096–0.53	2–3.7
Anthropogenic total	0.02	7.1–9.5	0.24	0.13–0.18	8.5–10.9
Natural total	9.2–15	1–2.6	1.4–2.7	0.26–1.3	12–22
S.H. total	9.2–15	9.1–13	1.6–2.9	0.39–1.5	20–33
Global total	15–26	78–111	6.1–10	2.2–6	102–154

[a] Includes biomass burning.

For DMS, the data indicate that 97% of the global flux of 15 to 26 Tg S/yr results from emissions from the ocean (Berresheim et al., 1995). Of this marine total, 61% is from the SH and 39% from the NH. The next largest contributor is split between wetland sulfur releases and those from anthropogenic/industrial emissions. By contrast, for SO_2 the global flux is largely defined by NH emissions (e.g., ~88%). This reflects the major contribution made from fossil fuel burning in the highly industrialized NH [for details see Spiro et al. (1992) and Hameed and Dignon (1992)]. Anthropogenic emissions from the SH make up still another 9% of the global total for SO_2 with volcanic emissions making up most of the remainder. This means that volcanic emissions define the second largest primary SO_2 global source but comprise, on average, only ~7% of the total. (Note, during years involving major eruptions, this source is substantially larger.) As noted earlier in the text, a very recent addition to the global inventory of SO_2 are emissions from ships. This source is currently estimated to be 2 to 4% of the total. In the case of other reduced sulfur (i.e., H_2S, CS_2, and OCS), the fluxes from the NH and SH are within a factor of 3 of each other and are made up of significant contributions from both natural and anthropogenic sources.

Overall, Table 1 clearly indicates that insofar as gross amounts of sulfur are concerned, fossil fuel combustion in combination with industrial emissions represent the single largest sulfur source in the NH. This is followed by nearly equal contributions from volcanoes and marine emissions. Still smaller emissions can be attributed to biomass burning and wetlands, and to direct release from plants and soils. By

contrast, in the SH ocean emissions of sulfur are nearly 2 times larger than those from fossil fuel combustion, the latter being followed by volcanic and ship emissions. [For a more in-depth survey see Bates et al. (1992b).]

A centrally important conclusion that can be extracted from Table 1 is that anthropogenic sources of sulfur have overtaken natural sources in the NH. For example, of the 101 Tg S/yr (on average) released in the NH, nearly 83 Tg S/yr (i.e., 82%) can be assigned to human activities. By contrast, for the SH only 37% can be similarly assigned. The fact that human-related activities are now overshadowing the natural sulfur cycle in the NH raises some serious questions as to what environmental price tag is being paid for such a transgression? Shifts in atmospheric acidity and atmospheric turbidity in the NH have now been documented, and new concerns are being voiced about the impact of elevated sulfur on regional weather patterns and in long-term climate changes [Charlson et al. (1992)]. Thus, sulfur, like other trace chemical substances in our environment, when present in too large amounts has the potential for creating deleterious consequences for humankind.

The global source strength of a sulfur species, in combination with its areal source distribution and atmospheric lifetime, typically define the species concentration level at a given location. This being true, given that several sulfur species have the same integrated global emission fluxes, the species having the more regionally focused source tends to generate the highest concentration levels. On the other hand, the longer the lifetime of a sulfur species, all other things being equal, the higher its concentration and the lower its variability. For example, as shown in Table 2, the very long lived species OCS (i.e., 2 to 4 years) has a global median concentration

TABLE 2 Observed Mixing Ratio of Atmospheric Sulfur Species

Sulfur Species	Background Marine Boundary Layer		Background Continental Boundary Layer		Polluted Continental Boundary Layer	
	Typical Range	Median	Typical Range	Median	Typical Range	Median
H_2S	2–30	10	5–150	60	80–810+	365
DMS	15–300	65	1–20	8	0–10?	< 5?
OCS	400–800	600	300–7000	550	300–1800	545
CS_2	1–35	10	15–50	30	65–370	190
SO_2	20–50	35	20–1000	500	150–6000+	1500
H_2S	7–13	9	1–7	6	IDTA	IDTA
DMS	0–20	2	IDTA	IDTA	IDTA	IDTA
OCS	1–8	4	< 3–18	7	IDTA	IDTA
CS_2	IDTA[a]	IDTA	IDTA	IDTA	IDTA	IDTA
SO_2[a]	10–80	30	60–260	100	IDTA	IDTA

[a] IDTA, Insufficient data to assess.

centered around 500 pptv and varies by no more than a factor of 2 on a regional and global scale. At the opposite end of this scale, Table 2 reveals that SO_2, which has both highly focused continental sources and a relatively short lifetime (i.e., 0.5 to 9 days), displays some of the largest gradients of any sulfur compound with concentrations ranging from 35 to 5000 pptv. Of particular significance is the gradient between background continental regions and remote marine areas where factors of nearly 15 are seen. By contrast, DMS, which has a somewhat similar lifetime to SO_2, typically shows far more modest boundary layer concentration gradients. This is in keeping with DMS having a far less focused source region. Interestingly, due to the combination of its short lifetime, efficiency of vertical mixing, and the absence of high altitude sources, DMS unlike SO_2 displays very significant altitudinal gradients. Similar arguments to those given for OCS, DMS, and SO_2 can be used to explain the concentration levels and gradients observed for other sulfur species.

3 TRANSFORMATIONS

As stated earlier, three of the major players in the atmospheric sulfur cycle are DMS, SO_2, and SO_4^{2-}. Reflecting this conclusion, the present section on transformations will primarily focus on the processes by which DMS and SO_2 undergo further oxidation to reach the final oxidation state of sulfur +6. Of special significance will be the +6 sulfur forms $H_2SO_4(g)$, SO_4^{2-}(non-sea-salt sulfate, NSS), and methane sulfonate (MS).

DMS Oxidation

Some of the earliest studies that attempted to define the oxidation products of DMS were those carried out by Niki et al. (1983), Hatakeyama et al. (1985), and Grosjean (1984) in the early 1980s. These studies can best be labeled as "chamber studies" in that they typically involved filling a large multi-liter vessel with air mixtures containing DMS, NO, and other trace species (i.e., HONO), and then activating the system with solar or artificial radiation to produce OH radicals. There have been many different versions of the chamber-type study [see reviews by Yin et al. (1990), Turnipseed and Ravishankara (1993), and Berresheim et al. (1995)], some starting with sulfur in the form of DMS while others have used intermediate oxidation products like DMSO. The early studies as well as those that have followed have been quite revealing in demonstrating that among the important oxidation products generated from DMS are SO_2 and MSA, with lesser amounts of DMSO and $DMSO_2$. In fact, all of these products have now been directly measured in the atmosphere using modern instrumental techniques.

Although qualitatively revealing, chamber studies have also had their limitations. This reflects the fact that the gas mixtures employed have been significantly different chemically than that which is typically found in a marine boundary layer (MBL) environment. In this case two of the more important species involved have been DMS itself and the radical scavenging species NO. Both have typically been present

in chambers studies at concentration levels several orders of magnitude higher than those found in the marine boundary layer. In addition, chamber studies have inherently been flawed due to their inescapably large surface-to-volume ratios (STVR). These have also been several orders of magnitude higher than those found in a marine environment (e.g., as aerosol surface area) and, thus, have led to greatly enhanced heterogeneous wall reactions. Since both concentration levels of DMS and NO as well as STVR factors impact on the DMS oxidation mechanism, not surprisingly, product distributions from individual chamber studies have been found to deviate significantly from study to study. They have thus left unanswered many of the quantitative details of the DMS product distribution within the MBL.

Among the more informative studies that have helped unravel aspects of the DMS oxidation mechanism have been those involving detailed laboratory kinetic investigations. These studies have focused on examining individual elementary reactions, the sum total of which, if available, would serve to define the overall DMS oxidation mechanism. One of the more pivotal of these was a study reported by Hynes et al. (1986). This study revealed that the reaction of OH with DMS proceeds not by a single reaction pathway but rather by two independent channels labeled kinetically as abstraction and addition, i.e., the reactions

$$CH_3SCH_3 + OH \Rightarrow CH_3S(OH)CH_3 \Rightarrow CH_2SCH_3 + H_2O \quad (1)$$

$$CH_3SCH_3 + OH \Rightarrow CH_3S(OH)CH_3 \quad (2a)$$

$$CH_3S(OH)CH_3 + O_2 \Rightarrow \text{products} \quad (2b)$$

As shown in Figure 3, the abstraction channel is nearly temperature independent; whereas, the addition reaction reveals a very significant negative temperature dependence. The crossover point for near equal contributions from both channels is seen as near 285 K. The early thinking on this mechanistic finding was that the OH abstraction channel was the channel that predominantly led to the formation of SO_2, while products such as DMSO, $DMSO_2$, and MSA were believed to be associated with the OH addition channel.

Evidence supporting the above position has included extensive field observations in which the stable end products MS and non-sea-salt sulfate (NSS) were measured and the value of their ratio then examined as a function of the ambient temperature. It was argued that SO_2 could be expected to undergo reasonably fast oxidation in the MBL via heterogeneous processes (see discussion under SO_2 Oxidation), thus forming NSS. On the other hand, MSA(g), formed from the addition channel, would be quickly scavenged by sea-salt aerosol to form MS. If indeed the above processes collectively define the mechanism by which both products are formed, as noted above the measured ratio of MS to NSS could be expected to provide a good chemical reflection of the average temperature at which the DMS oxidation occurred. In fact, extensive field measurements that have evaluated this ratio over a range of latitudes and altitudes have shown that the lowest values (e.g., 0.07) occur at tropical latitudes and that some of the highest values (e.g., ≥ 0.34) tend to be found at much higher latitudes (e.g., Berresheim, 1987; Savoie and Prospero, 1989; Bates et al., 1992a). However, a more limited but still quite significant number of

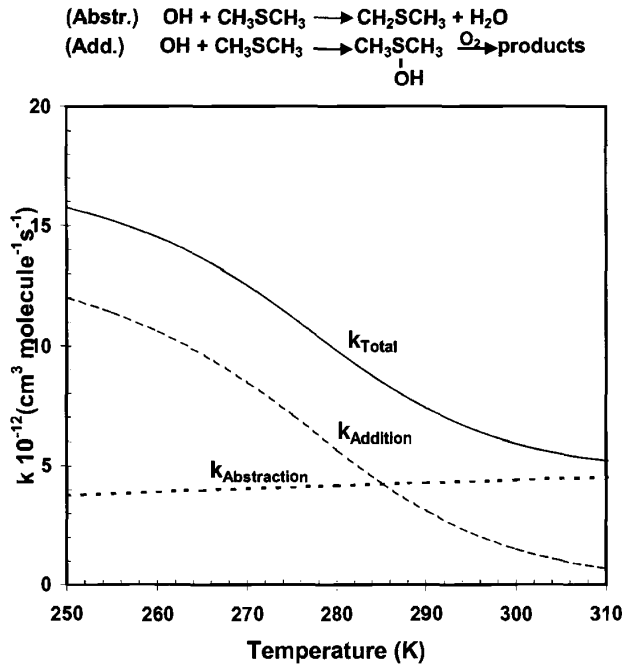


Figure 3 Temperature dependence of the rate coefficients for the OH/DMS addition and abstraction reaction channels as well as the total k value (modified from Berresheim et al., 1995).

observations have also been reported that do not follow this simple trend (e.g., Berresheim et al., 1995; Davis et al., 1998). Since these observations appear to be equally valid, they most likely point to a DMS oxidation mechanism that is more complex than originally thought. (Note, the potential importance of the MS/NSS ratio as defined by DMS oxidation rests in the fact that this ratio, if well understood, could be used to apportion the DMS contribution to total NSS. Perhaps, more importantly, it could be used as an indicator of the temperature environment under which the DMS oxidation process took place.)

One rendition of the DMS oxidation mechanism that reflects the thinking that the overall process is actually quite complex is that shown in Figure 4. This mechanism (shown here in abbreviated form) has folded in the most recent results from both field studies as well as laboratory kinetic investigations. Quite significant is the clear indication that not only is the OH abstraction channel a source of SO_2 but that the addition branch, in several different steps, also can form SO_2 as a product. Equally important, the stable product MS is shown as a product of both addition and abstraction channels. Its production efficiency is shown as being even further convoluted as

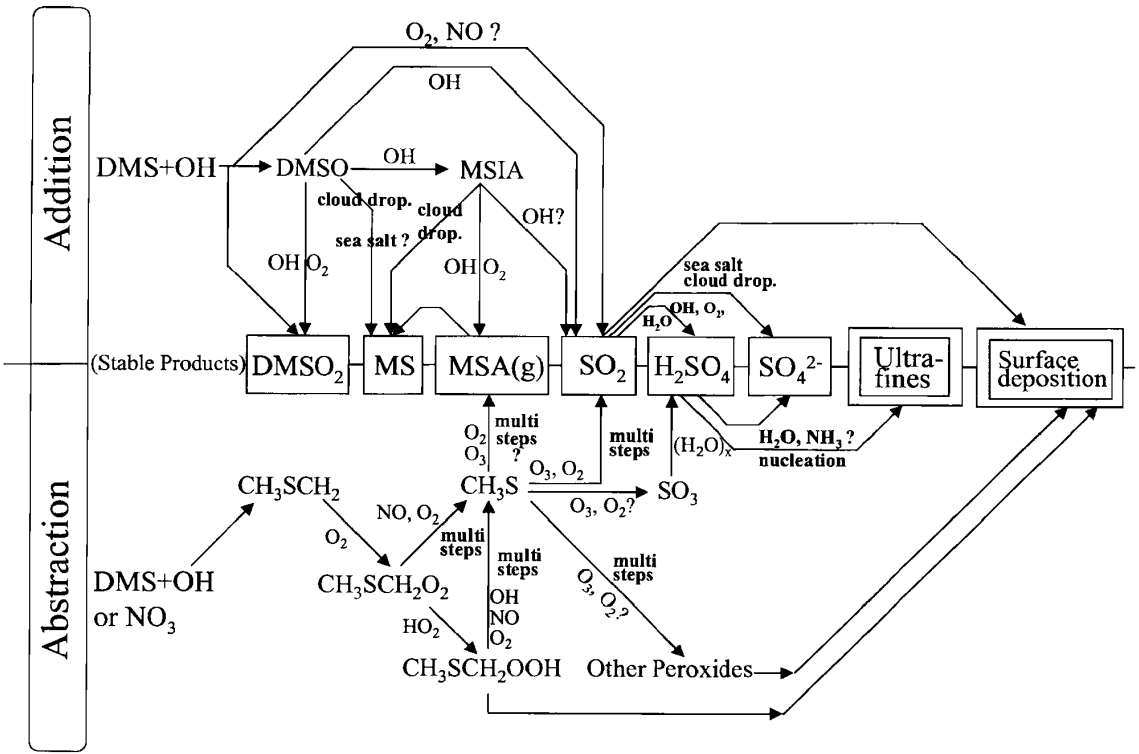


Figure 4 Abbreviated DMS oxidation scheme (modified from Davis et al., 1999).

a result of it being formed through competing gas and heterogeneous processes involving the intermediates DMSO and MSIA. Although the mechanism shown is still speculative (e.g., many of the elementary reactions have not yet been fully characterized), recent sulfur field studies, covering a wide range of latitudes, have provided evidence that strongly supports key aspects of this mechanism.

Unlike some of the earliest sulfur field studies, more recent investigations have reported a significant coupling between DMS and SO_2. Given that both DMS and SO_2 typically have MBL lifetimes of 0.5 to 2 days, if DMS is a significant source of SO_2, one would expect that these two species should be anticorrelated when measured at a rate significantly shorter than their respective lifetimes. Alternatively, if measured at a time resolution significantly longer than their respective lifetimes, one would expect a positive correlation. Thus, depending on the sampling rate, the appearance of either a correlation or anticorrelation in field data would signal there being a significant oxidative pathway from DMS to SO_2. As cited above, in the earliest studies for which simultaneous measurements of DMS and SO_2 were recorded, no relation between these two sulfur species was found. However, with improvements in instrumentation and the more judicious selection of field sites (e.g., free of anthropogenic pollution sources), a quite different picture is now emerging. Among the more significant studies has been that reported by Bandy et al. (1996), which took place in 1994 at Christmas Island. Located at 2°N in the middle of the Pacific Ocean, Christmas Island defines an ideal setting for studying DMS oxidation chemistry. Situated near the middle of the equatorial upwelling, it experiences near year-round elevated levels of DMS with no evidence of significant other sources of SO_2. In addition, being located well within the strong trade wind regime, it typically experiences stable meteorological conditions for several days at a time. Finally, due to the high solar flux and water vapor levels present, it also defines an environment where very high levels of the critical boundary layer oxidizing agent OH can be found. Reflecting these optimum conditions, Bandy and co-workers (1966) reported the first convincing field data showing a strong diel relation between DMS and SO_2. These investigator's high temporal resolution data were recorded over a 9-day time period and revealed a clear and convincing anticorrelation between these two sulfur species. The estimated DMS to SO_2 conversion efficiency was reported as $62 \pm 6\%$.

In an airborne follow-up study at Christmas Island in 1996 [part of the National Aeronautics and Space Administration's (NASA's) GTE PEM–Tropics A program, Hoell et al. (1999)], the sulfur database reported was even more revealing. During this investigation direct observations were recorded of both DMS and SO_2 as well as the oxidizing agent OH. Equally significant was the availability in this new study of meteorological and chemical data as a function of altitude. As shown in Figure 5a, the profiles for DMS, SO_2, and OH make for a very convincing case that DMS is a major source of SO_2 and that DMS oxidation predominantly occurs via OH radicals. An analysis of these new data by Davis et al. (1999) resulted in an overall DMS to SO_2 conversion efficiency of $72 \pm 22\%$, well within the range reported by Bandy et al. (1996). Given that the abstraction channel at the temperatures of Christmas Island (298 K) represented 70% of the total OH/DMS reaction rate, together with the conservative estimate that at least 8% of the product yield from DMS forms species

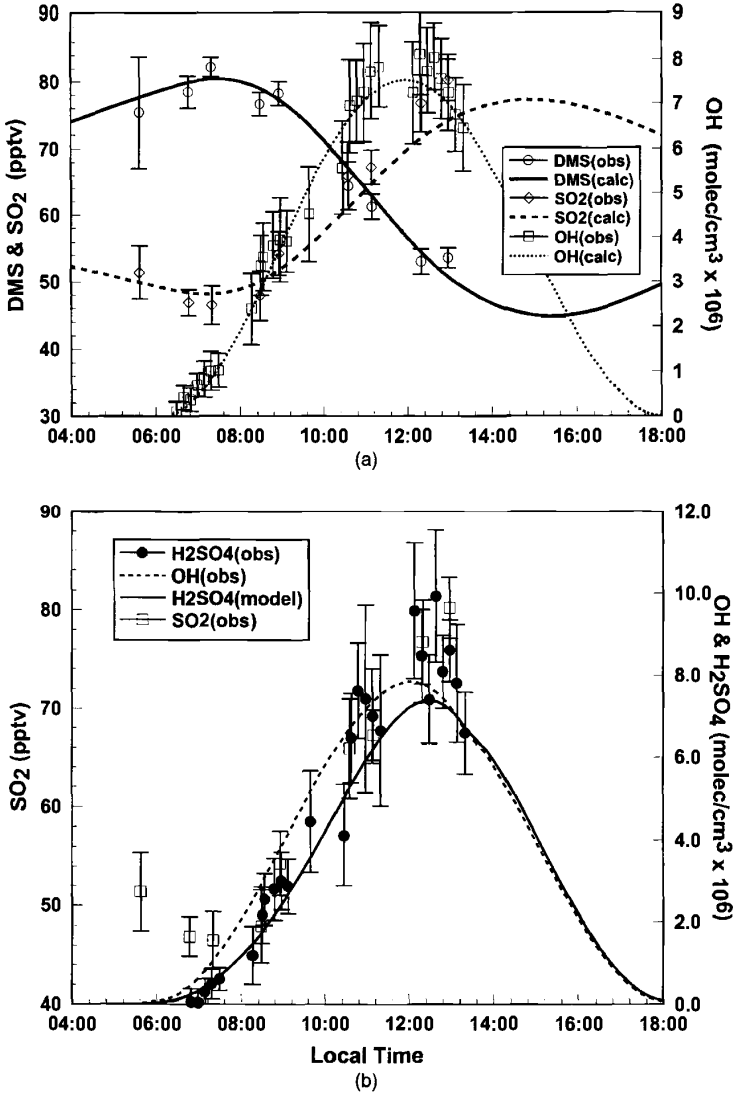


Figure 5 Analysis of airborne sulfur data collected during NASA's PEM–Tropics A program: (*a*) observed and model profiles for DMS, SO_2, and OH for tropical boundary layer conditions; (*b*) plot showing SO_2 conversion to H_2SO_4 via reaction with OH (see e.g. Davis et al, 1999). [Note: DMS and SO_2 observational data are those recorded by Thornton et al. (1999). The OH and H_2SO_4 observational data are those recorded by Mauldin et al. (1999a,b)].

other than SO_2, one can estimate that the likely range for SO_2 formation from the abstraction channel is between 0.6 and 0.9. This suggests that the contribution of SO_2 from the addition channel is probably not much lower than 0.4; however, the issue of SO_2 contributions from the addition channel is more fully and better explored in the text below based on field studies conducted at much lower ambient temperatures.

Shifting to the recent National Science Foundation/National Oceanic and Atmospheric Administration (NSF/NOAA) program ACE-1 [Aerosol Characterization Experiment, Bates et al. (1998)], the opportunity again presented itself to examine DMS oxidation chemistry under remote conditions, but with the field site being defined as the Southern Ocean, just to the south of Tasmania. Thus, it provided an environment having much lower ambient temperatures. Recall that because of the strong negative temperature dependence of the OH/DMS addition channel, the addition channel becomes more prominent under these conditions. In fact, based on the average temperature recorded on the ACE-1 aircraft sampling platform (i.e., 280 K), both channels are estimated to be of near equal importance. Davis (unpublished results) in his analysis of the resulting DMS and SO_2 data has estimated that the overall DMS-to-SO_2 conversion efficiency to be still quite high, e.g., 0.7 to 0.9. Thus, to be consistent with the measured SO_2 levels and the previously cited average efficiency for SO_2 production from the abstraction channel of 0.75, he assigned a range of 0.7 to 0.9 to the addition channel. Although seemly quite high, the latter range is seen as being consistent with the previously discussed SO_2 tropical analysis.

In yet another study at still higher latitudes (i.e., 66°S), the land-based NSF SCATE Antarctic Program (Berresheim and Eisele, 1998), the average temperature recorded was only 273 K. In this case the addition channel is estimated to be 65% of the total OH/DMS kinetic rate. Although SO_2 measurements were not made during this campaign, direct observations of OH, DMS, and H_2SO_4 suggest that the overall SO_2 conversion efficiency required to support the observed H_2SO_4 levels would need to be well above 60%. Thus, these data again point toward an SO_2 efficiency for the addition channel of 50% or higher (Davis et al., 1998).

Quite interestingly, in all of the analyses cited above, involving high conversion efficiencies of DMS to SO_2, the measurements being used have been those reported by Bandy and co-workers (1996). The latter group pioneered the use of the isotopic-dilution gas-chromatographic mass-spectrometric technique (IDGCMS) for both field measurements of DMS and SO_2 (Bandy et al., 1993). On the other hand, in two recent ship-based studies, one in the tropical South Pacific [MAGE (Marine Aerosol and Gases Experiment), Yvon et al. (1996)], the other at high latitudes as part of the ACE-1 program (De Bruyn et al., 1998), a quite different technique was employed in the measurement of SO_2. In both of these cases SO_2 measurements were made using the aqueous-phase fluorescence method (Saltzman et al., 1993). In both studies, evidence was found of diel trends in DMS and SO_2, further confirming the important role of photochemical oxidation of DMS. The reported overall conversion efficiencies for DMS to SO_2, however, were estimated to be significantly different from those cited above. In each of the ship studies, the conversion effi-

ciencies ranged from 30 to 50%, a factor of 1.5 to 2 lower than those already cited. It seems unlikely at this juncture, even though the latter studies were conducted on different platforms, that the results would differ by this amount. Both the average temperature and levels of other critical chemical species appear to have been similar for the tropical studies and the high latitude investigations. Whether the reported difference is due to SO_2 measurement difficulties or to yet unknown factors cannot be determined at this time.

Evidence that oxidizing agents (i.e., NO_3 and Cl) other than OH are important in converting DMS to SO_2 has been primarily based on results from laboratory kinetic studies [see, e.g., Berresheim et al. (1995) and references therein]. In combination with model estimated levels of NO_3 and Cl, the tentative conclusion has been reached that only the NO_3 mechanism is significant and that for most marine areas even this oxidant is unimportant. The exception would be for highly populated coastal influenced regions where major sources of NO_x could be expected. As related to the importance of Cl atom oxidation DMS, even though some recent sulfur field data (e.g., MAGE study) suggest that the impact from Cl might be significant, other results reveal a different picture. For example, the results from the previously discussed field studies at Christmas Island as well as those in the Southern Ocean suggest that Cl atom DMS oxidation is less than 15%. An independent study by Singh et al. (1996), in which C_2Cl_4 budget arguments were used to evaluate the significance of boundary layer Cl atom oxidation, also resulted in a similar conclusion, namely, that the latter chemistry is of negligible importance in remote marine regions.

In addition to the pivotal question related to the oxidative conversion efficiency of DMS to SO_2, significant other DMS oxidation issues also continue to be the subject of continuing research. These include identifying the DMS oxidation intermediate(s) responsible for gas-phase $H_2SO_4(g)$ formation, the elucidation of the pathways by which DMSO, $DMSO_2$, MSA(g), and MS are formed, and the identification of the factors controlling the MS/NSS ratio. In the text that follows these DMS issues are further explored in the context of both recent laboratory kinetic investigations as well as the results from recent marine sulfur field studies.

As related to $H_2SO_4(g)$ formation, a review of the literature points to two major reaction sequences as being of potential importance. These include (3a) to (3d) and (4a) and (3d), e.g.,

$$\text{DMS} + \text{OH multisteps} \Rightarrow \text{SO}_2 + \text{other products} \quad (3a)$$

$$\text{SO}_2 + \text{OH} + \text{M} \Rightarrow \text{HSO}_3 + \text{M} \quad (3b)$$

$$\text{HSO}_3 + \text{O}_2 \Rightarrow \text{SO}_3 + \text{HO}_2 \quad (3c)$$

$$\text{SO}_3 + (\text{H}_2\text{O})x \Rightarrow \text{H}_2\text{SO}_4(\text{g}) + \text{H}_2\text{O} \quad (3d)$$

$$\text{DMS} + \text{OH multisteps} \Rightarrow \text{SO}_3 + \text{other products} \quad (4a)$$

$$\text{SO}_3 + (\text{H}_2\text{O})x \Rightarrow \text{H}_2\text{SO}_4(\text{g}) + \text{H}_2\text{O} \quad (3d)$$

Two of the most revealing recent field studies that have examined this issue are the previously discussed PEM–Tropics A Christmas Island airborne study and the

ground-based SCATE study in Antarctica. Recall, that during the PEM–Tropics A field investigation direct observations of DMS, OH, SO_2, H_2SO_4, and total aerosol surface area were simultaneously recorded. As shown in Figure 5*b*, using the known rate coefficients for processes (3a) to (3d), together with recently measured aerosol sticking coefficients for H_2SO_4, Davis et al. (1999) concluded that the observed profile for H_2SO_4(g) could be convincingly explained in terms of the observed diel profiles for SO_2 and OH. Since, as previously discussed, the observed SO_2 profile from this field study was also explicable in terms of OH/DMS oxidation, these new results are consistent with the idea that SO_2 is the critical DMS intermediate leading to gas-phase H_2SO_4 formation. Taking a different approach, Jefferson et al. (1998), not having direct observations of SO_2, used the observations of DMS, H_2SO_4, OH, and total aerosol surface area from the SCATE program to evaluate the two quantities k_{OH} [OH][DMS] and $k_{surf}[H_2SO_4]$. It was argued that if the direct formation of SO_3 from DMS were important, given the reasonably short lifetimes for both H_2SO_4 and SO_3 in the Antarctic environment (i.e., <1 h), one would expect a significant correlation between these two quantities. In fact, the R^2 value was less than 0.2, indicating no relationship. Although still lacking the finality that comes from having a comprehensive set of elementary rate constants for each step in a mechanism, the collective results cited above strongly suggest that SO_2, not SO_3, is the dominant intermediate from the oxidation of DMS that leads to the gas-phase formation of H_2SO_4.

Field observations bearing on the mechanistic details surrounding the formation of DMSO and $DMSO_2$ from DMS oxidation have been limited in number and conflicting in their results. They include results from a sulfur field study near the Washington coast, the previously discussed Antarctic SCATE program, and finally the 1994 Christmas Island study. During the SCATE program, there were ~6 days of near continuous recording of DMSO and $DMSO_2$ (Berresheim et al., 1998). However, in the analysis of these data Davis et al. (1998) could find only 1 day out of the 6 sampled in which it appeared that both DMSO and $DMSO_2$ levels were controlled by local photochemical production. For all other days DMSO and $DMSO_2$ were shown to be controlled by transport processes, wherein large quantities of ocean-released DMS were initially carried aloft into the lower free troposphere, oxidized, and then returned in the form of intermediate as well as +6 oxidation state sulfur. But, on January 19, 1994, it appears that there was a significant break in this cycle in that background levels of both DMSO and $DMSO_2$ were found to be a factor of 10 lower than during the other 5 sampling days. Only on this day, was there any evidence of a diurnal profile for DMSO that tracked the measured ultraviolet (UV) solar irradiance. From their analysis of these data, Davis and co-workers (1998) estimated that the DMSO formation efficiency from the OH/DMS addition channel could range from 0.5 to 1.0, a value well within the limiting value assigned to this branching ratio based on two independent laboratory kinetic investigations. The $DMSO_2$ data, although considerably more noisy, were found to be most consistent with a branching efficiency for DMSO/OH to $DMSO_2$ of ~0.3.

In the field study near the Washington coast (Berresheim et al., 1993), 2 to 3 days of DMS and DMSO data were collected, but with very little ancillary data to facil-

itate defining the photochemical environment for this investigation. On April 14, however, sunny conditions prevailed nearly all day, and DMS and DMSO were sampled while air was advected in from the Pacific Ocean. For this specific case Berresheim et al. (1995) were able to estimate a branching ratio for the DMS/DMSO addition channel of ~ 0.5 but with a large uncertainty. This result may again be viewed as in good agreement with the above-cited results by Davis et al. (1998), but both have large uncertainties associated with them. In the 1994 Christmas Island study the measured levels of DMSO were found to be incredibly large relative to the median values of DMS observed, i.e., median DMSO 25 pptv, median DMS 200 pptv. Chen et al. (2000) in their analysis of this data quickly concluded that the two observations were totally irreconcilable in terms of any known DMS oxidation mechanism. This led them to put forward two possible hypotheses: (1) There were possibly unknown difficulties in the measurements of DMSO or (2) yet unknown sources of DMSO may exist. Still more recently NASA's PEM Tropics B Field study, direct airborne measurements of DMS, DMSO, and OH from sunrise to 1pm local time indicated that the highest levels of DMSO were at or near sunrise. Concentrations were found to decrease throughout the remainder of the measurement period (Nowak et al., 2001). The authors have pointed out that the temporal behaviour of DMSO is totally contrary to that expected if DMSO was formed only from the reaction of DMS with OH. Thus, these new data also suggest an additional source of DMSO, one that operates at night as well as possibly during daylight hours. Quite clearly if the latter hypothesis is correct, Table 1, as related to global sources of sulfur, would require further modification.

The issue of how efficiently DMSO might be formed through the OH/DMS addition channel raises the equally important question: Does the further oxidation of this species provide an effective pathway for formation of MSA(g)? This +6 oxidation state sulfur compound is shown in Figure 4 being formed from the oxidation of methane sulfinic acid (MSIA), another intermediate from the OH/DMS addition channel. Recent laboratory kinetic data (Hynes et al., 1996; Urbanski et al., 1998) would seem to support this notion in that they found the OH/DMSO reaction to be very fast and that the reaction appears to lead to near unity yields of the CH_3 radical. The product CH_3 radical is one that would be expected if the initial adduct formed from the reaction of OH with DMSO broke apart to form MSIA. Even so, in both of the airborne field studies previously cited, as well as during project SCATE, direct observations of gas-phase MSA have revealed a very low production efficiency for this species, e.g., typically $\leq 1\%$. The latter result is significant in that it translates to our explaining no more than 2 to 5% of the observed MS aerosol loading from the condensation of MSA(g). This means that both under tropical as well as the low-temperature conditions of the Antarctic, gas-phase production of MSA is not the major source of MS (the latter being a frequently cited measurement in much of the older literature involving DMS field studies).

The above MSA(g) results again focus our attention on the question touched on earlier in the text: How well do we really understand the factors controlling the value of the much cited MS/NSS ratio? Recall earlier (bottom of page 132) in the text we discussed the fact that some observations have shown a trend of increasing values in

this ratio with decreasing temperatures (i.e., increasing latitude) but that notable exceptions had also been seen in this trend. One recent explanation for both the observed low yield of MSA(g) and yet significant yields of MS has been that proposed by Jefferson et al. (1998). To explain the MSA(g)/MS SCATE results, these investigators proposed that the rather high median levels of 1 to 2 pptv of DMSO observed on the Palmer Peninsula (see earlier discussion in text related to DMSO formation at Palmer) could only be accounted for if heterogeneous DMSO reactions were occurring on sea salt aerosols. This hypothesis is supported by the fact that in several independent aqueous phase kinetic studies, DMSO has been shown to react in the aqueous phase (in the presence of oxidizing agents such as OH radicals) to form MSIA. This species was observed to subsequently undergo further oxidation to yield MS. Davis et al. (1999) came to a similar conclusion when analyzing the PEM–Tropics A data; however, these investigators noted that both gas-phase DMSO and MSIA(g) would be equally good candidates as a heterogeneous source of MS. In yet another laboratory kinetic study, Lee and Zhou (1994) examined the aqueous-phase reaction of DMS with O_3 as a possible source of aerosol-phase oxidized sulfur. What they found was that because of the very low Henry's law constants for both DMS and O_3, the probability of this aqueous process is rather unfavorable. However, under the most favorable conditions involving heavy clouds and relatively high O_3 levels, it could prove to be a significant source of oxidized sulfur. Collectively, the above findings would seem to suggest that the formation of MS and possibly other sulfur species must be viewed in the context of both gas-phase and heterogeneous chemistry in the atmosphere.

The implications of the above findings are quite significant in that they clearly point to the possibility that the MS/NSS ratio depends not only on the temperature of the environment where DMS oxidation occurs, but is equally, if not more, influenced by the nature of the aerosol environment, e.g., sea salt loading, cloud density, and liquid water content. For very low aerosol loadings a substantial fraction of the DMSO and MSIA(g) from the OH/DMS addition channel would most likely react with OH in the gas phase to produce SO_2. Most of this SO_2 would subsequently be converted into NSS. On the other hand, for very high aerosol loadings nearly all DMSO and MSIA would likely be scavenged, producing MS as a final product. Thus, for a given sampling location where the aerosol loading might vary from day to day, one could expect to find a range of MS/NSS values. In this context, one of the most stable environments in which MS/NSS values would be rather constant would be that defined by the tropical marine BL. Here the abstraction branch would strongly dominate DMS oxidation and the sea-salt aerosol loading would remain both relatively high and reasonably constant. Indeed, some of the most consistent values for the MS/NSS ratio have been those measured in the tropics (e.g., Saltzman et al., 1986; Savoie and Prospero, 1989; Berresheim et al., 1995). However, in spite of what appears to be a reasonably well documented environment, one should not lose sight of our earlier discussion that hinted at the strong possibility that there may be a significant and yet unidentified source of DMSO. If so, considerable rethinking of tropical marine sulfur chemistry may be necessary.

Thus far our DMS discussions have been primarily focused on the marine boundary layer (MBL). It may be asked, therefore, how dramatically does this picture change if the oxidation of DMS were to occur in the free troposphere (e.g., above 2 km)? In fact, as hinted at in our earlier discussions of the temperature dependence of the OH/DMS reaction, quite significant changes can occur. In the free troposphere three major physical changes occur in the environment: the temperature drops (e.g., 6.5°C/km), the pressure drops (i.e., exponentially), and the average aerosol surface area drops by at least one order of magnitude. It is the first and third of these shifts that potentially could have the most significant impact on DMS oxidation chemistry. Recall, that the OH/DMS addition channel has the strongest dependence on temperature (increasing with decreasing temperature), and therefore this channel becomes the dominant one with increasing altitude. On the other hand, laboratory studies suggest that this channel probably also has the greatest diversity in oxidation products. Equally important, the oxidation product distribution from this channel appears to have the greatest dependence on aerosol surface area, i.e., heterogeneous reactions. Thus, speculating on the net effect of these factors might point toward enhanced levels of both $DMSO_2$ and MSA(g). It could also mean that a much larger fraction of the DMSO and MSIA would be oxidized via OH, leading to higher yields of SO_2 or new products like sulfurous acid (H_2SO_3) from this channel. This sequence of reactions, in turn, could lead to the higher yields of H_2SO_4(g) which under the cold temperatures of the upper troposphere could form the basis for new aerosol particle formation as suggested by Clarke (1993). Still another interesting result from this high-altitude DMS chemistry would be its impact on the MS/NSS ratio. For example, with an enhancement in the yield of SO_2 from the addition channel, the value of this ratio might remain reasonably low even though the temperature at which the oxidation occurred was quite low. Suffice it to say, both new laboratory kinetic studies as well as field observations will be required to actually quantify this chemistry.

SO_2 Oxidation

As both a primary source species (e.g., combustion and volcanoes) and as one of the major products from DMS oxidation, the atmospheric fate of SO_2 represents a major component of the atmospheric cycling of sulfur. It is estimated that between 40 and 60% of this SO_2 is directly deposited to either land or ocean surface areas (Berresheim et al., 1995). The remainder is believed converted into sulfur +6, although the detailed mechanisms by which this final oxidation state is reached continues to be the focus of ongoing research. What is now reasonably clear is that there are at least two general pathways by which this is achieved: one involving gas-phase chemistry, the other involving heterogeneous reactions. The gas-phase process is now relatively well understood, involving the ubiquitous oxidizing agent OH, e.g., reactions (3b) to (3d). The first two steps were reasonably well established by the mid 1980s (Finlayson-Pitts and Pitts, 1986); however, only recently have the details of step (3d) been established (Lovejoy et al., 1996). This process has now been shown to involve a quadratic dependence on H_2O. But, considering the amount of H_2O in the atmo-

sphere, this step is rarely if ever the rate-limiting step. In virtually all cases step (3b) is rate limiting.

As related to the gas-phase oxidation of SO_2, the importance of this process must be viewed both from the perspective of converting bulk atmospheric SO_2 to sulfate and from the point of view of its role as a major source of gas-phase H_2SO_4. Current evidence suggests that the gas-phase oxidation of SO_2 is probably no greater than 20% of the total, and in the final analysis it could be no more than 5 to 10% (Lelieveld and Heintzenberg, 1992). On the other hand, the gas-phase production of H_2SO_4 now appears to represent a critical step in the formation of new particles (via heterogeneous nucleation) that ultimately leads to cloud formation (e.g., Kreidenweis and Seinfeld, 1988). Thus, in the absence of this source, it would be difficult to explain how the atmosphere resupplies itself with CCN. CCN are routinely removed by both wet and dry deposition. This suggests then that the gas-phase oxidation of SO_2 is of primary importance in the atmosphere defining the strong link between sulfur emissions and climate effects.

Of the 80 to 90% of the SO_2 that is oxidized by non-gas-phase pathways, both heterogeneous reactions involving cloud droplets as well as sea salt aerosols are considered important. [For highly industrialized regions the influence of soot particles, trace metals such as Fe(+3), Mn(+2), and Cu(+2), and organic carbon reactions must also be included.] In remote marine areas, the heterogeneous oxidation process typically involves several steps. The first of these involves the critical equilibria shown in (5a) and (5b):

$$SO_2(g) + H_2O \rightleftharpoons SO_2(aq) \tag{5a}$$

$$SO_2(g) \rightleftharpoons HSO_3^- \tag{5b}$$

$$HSO_3^- \rightleftharpoons SO_3^{2-} + H^+ \tag{5c}$$

The presence of these equilibrium reactions means that the dominant form of sulfur +4, in the bulk aqueous phase, depends very much on the acidity of the aerosol species. For the most typical range of acidity in the troposphere, the dominant form of sulfur is the bisulfite ion (HSO_3^-). However, because of shifts in the levels of the individual forms of sulfur with changing pH, as well as the dependence of the reaction coefficients on pH, the most important aqueous-phase pathway for oxidation of sulfur +4 can be a strong function of pH, and therefore on the total amount of sulfur converted (e.g., Martin, 1984). This point is illustrated in Figure 6. Here it can be seen that for pH values above 5, the oxidation by O_3 represents the dominant pathway; whereas for pH values less than 5, oxidation via H_2O_2 becomes the major source of sulfur +6. Other investigators (e.g., Chameides and Stelson, 1992; Sievering et al., 1992) have proposed that the sensitivity of the aqueous-phase oxidation of sulfur +4 to percent sulfur converted might be much smaller than originally thought. It has been suggested that this would be particularly true when the aerosol species is sea salt. The above group of investigators have argued that sea salt contains a natural buffering capacity involving the bicarbonate/carbonate system. Thus, seawater aerosol might be able to sustain a high rate of conversion of +4 sulfur to +6 through the O_3 oxidative pathway for extended periods of time.

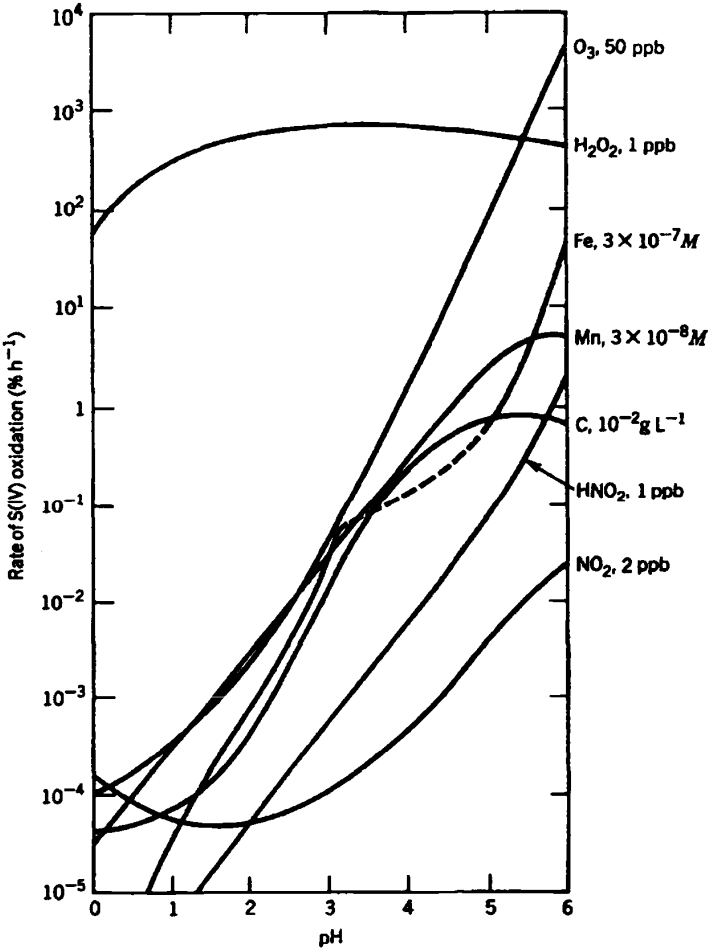


Figure 6 Estimated rates of oxidation of S(IV) in solution and on carbon surfaces as a function of pH (taken from Martin, 1984).

Although the above discussion might be viewed as downplaying the overall role of atmospheric photochemistry in the conversion of sulfur +4 to the +6 state, this clearly is not the case. For example, not only are there potentially other aqueous-phase reactions driven by scavenged gas-phase radicals such as HO_2 (e.g., Chameides and Davis, 1982); but there is also the fact that the critical heterogeneous oxidizing agents are typically O_3 and H_2O_2. Both of these species are themselves predominantly generated in the gas phase via photochemical processes. Thus, both

the SO_2 oxidation by OH and that by heterogeneous pathways must be viewed as significantly influenced by local and/or regional photochemistry.

4 GLOBAL DISTRIBUTIONS OF SO_2 AND SULFATE

In Section 1 the point was made that of the stable atmospheric forms of sulfur, SO_2 and SO_4^{2-} were among the more important species that need to be well understood. In this context, a key characteristic that needs to be examined for both species is its atmospheric lifetime. Both tend to be long enough to permit their being transported over near hemispheric scales. Thus, their global distributions represent an important

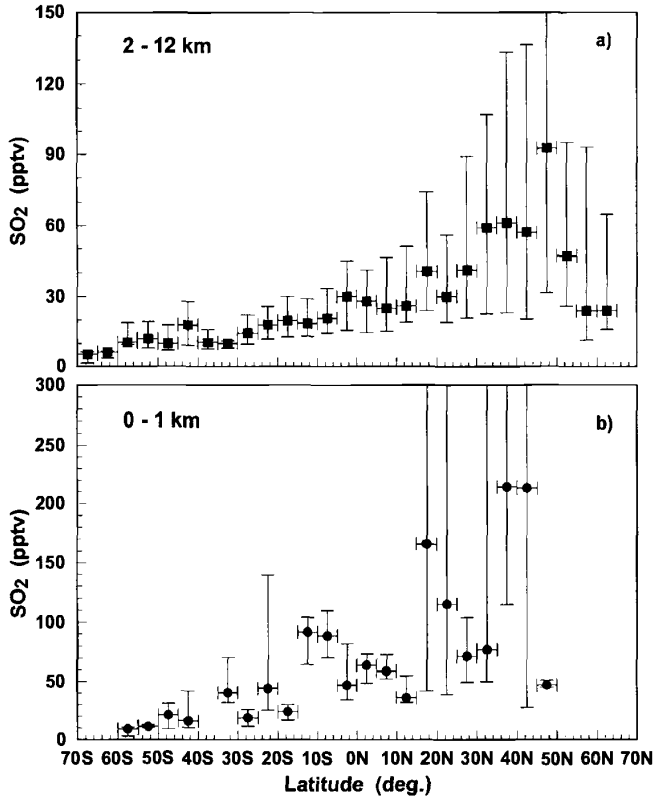


Figure 7 Latitudinal distributions of SO_2 for the altitude ranges of: (*a*) 0 to 1 km and (*b*) 2–12 km (modified from Thornton et al., 1999). The data in these plots have been derived from individual observations that have been binned every 5° of latitude. Vertical bars on each 5° latitude bin signify the one sigma variability in the average value for each bin.

indicator of humankind's influence on the natural sulfur cycle. At present the database for both species is still quite limited with vertically resolved data now being available for no more than 20% of the global atmosphere. Most of the latter data is also limited to no more than one or two seasons of the year with the geographical coverage being confined largely to the North and South Pacific Oceans. Representative of these data are the latitudinal plots of SO_2 shown in Figures 7*a* and 7*b*. For clarification purposes, these data have been binned for the altitude ranges of 0 to 1 and 2 to 12 km. Several points made earlier in the text, concerning anthropogenic effects, are clearly revealed in these plots. For example, the Northern Hemisphere is seen as having an average mixing ratio for SO_2 that is nearly five times higher than that for the Southern Hemisphere. Equally significant is what appears to be direct evidence for the focused release of SO_2 in the highly industrialized midlatitude region of 30 to 55°N.

Thus, given the limitations of current field data, one must turn to models to explore in greater depth the global atmospheric picture of sulfur. In this case the goal is that of gaining further insight into the distributions and variations in sulfur compounds and the processes that regulate their concentration levels. Several global-scale chemistry transport models have been developed in the past 7 years for just this purpose. These models endeavor to place available input sulfur data, as related to sources, sinks, and concentration levels, into a comprehensive global sulfur cycle (e.g., Langner and Rodhe, 1991; Pham et al., 1995; Feichter et al., 1996; Chin et al., 1996, 2000; Chuang et al., 1997; Koch et al., 1999; Barth et al., 2000). All include tropospheric sulfate and its major precursors (i.e., DMS and SO_2) and contain modules designed to handle anthropogenic and natural emissions, chemical transformations, advection/convection, and dry and wet deposition. Illustrative of the output from these models, we show in Figure 8*a* and 8*b* the annually averaged global surface-air distributions for SO_2 and sulfate based on results from Chin et al.'s (2000) model. This model includes sulfur from fossil fuel and biofuel combustion, shipping and aircraft emissions, biomass burning, volcanoes, and biogenic sources. Here it can be seen that the maximum SO_2 concentrations are clearly located at latitudes between 30 and 75°N, corresponding to the major industrial source regions of eastern United States, Europe, and eastern Asia. The levels of SO_2 are seen ranging from 1 to over 10 ppb. Interestingly, significant surface SO_2 concentrations are also shown to be present over southern Africa and Chile, largely reflecting ore smelting operations. The distribution of surface-air sulfate over the continents is found to be very similar to that for SO_2, although the gradients are clearly smaller. These observations reflect the fact that sulfate is the primary product of SO_2 oxidation and that transport as well as dry and wet deposition represent major losses for SO_2 (see discussion later in chapter). The model results also reveal that sulfur concentrations in the Arctic and near coastal regions in the Northern Hemisphere tend to be heavily influenced by anthropogenic releases. Returning to the field data shown in Figure 7*a*, the model values found at these near continental locations appear to be substantial larger than those reported in the latitudinal plots taken from the data of Thornton et al. (1999). Recall, however, that most of these data were recorded over remote regions of the Pacific. The sulfate distribution (i.e.,

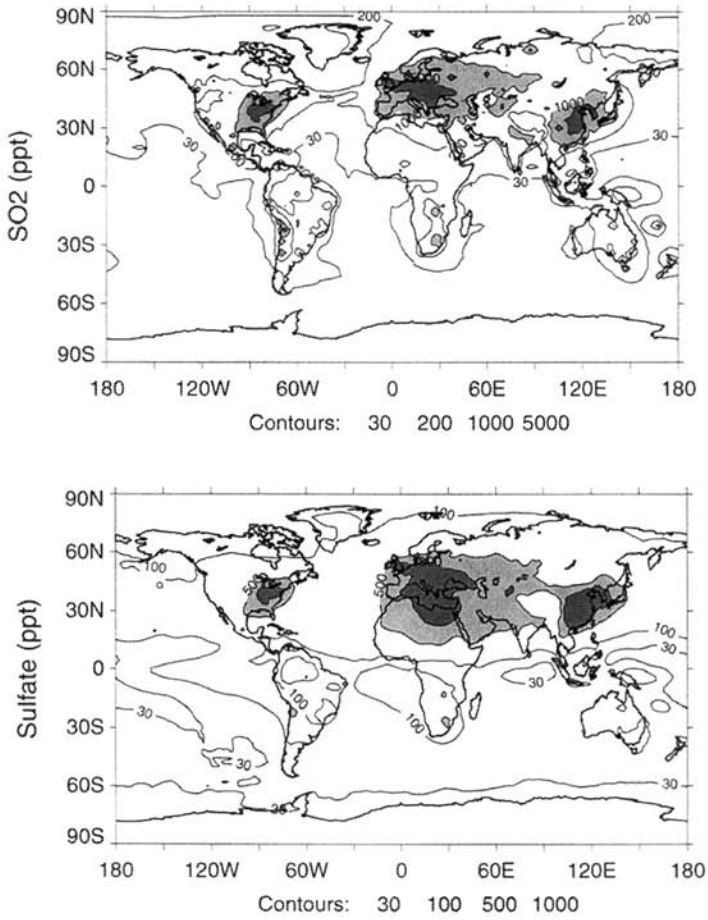


Figure 8 Global sulfur mixing ratios in the lowest 500 m as derived from the global chemistry transport model of Chin et al. (1999): (*a*) SO_2 and (*b*) SO_4^{2-}.

Figure 8*b*) which also is shown falling off like SO_2 as one moves from continental regions to the open ocean, is similarly in good agreement with observational data when one considers the geographical location of the surface sampling sites [see, e.g., the data of Savoie et al. (1989) over the North Atlantic, that of Savoie and Prospero (1989) over the North Pacific, and that in the Arctic reported by Barrie et al. (1989)].

As seen in the SO_2 data of Thornton et al. (1999) and in the model results shown in Figure 9*a* and 9*b*, the impact from anthropogenic emissions of sulfur is significantly attenuated at altitudes well above the boundary layer. This is due both to a large fraction of the combustion-based SO_2 and sulfate being deposited within the continental boundary layer and also to the more efficient dispersion of these species once at higher altitudes. Chin and Jacob (1996) have estimated that about 40% of the Northern Hemisphere industrial source of SO_2, is transported out as SO_2, or sulfate

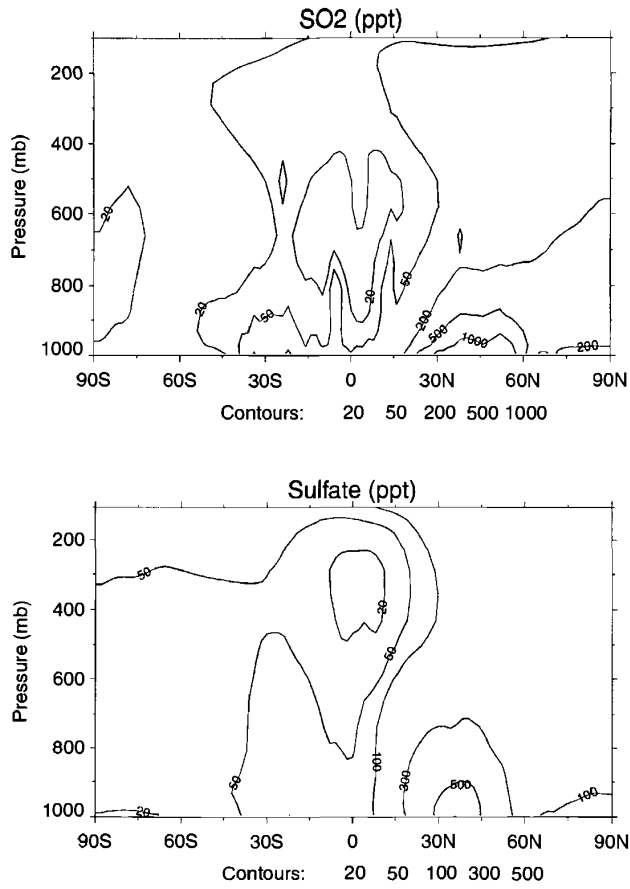


Figure 9 Global altitudinal and latitudinal distribution of the sulfur mixing ratio based on the global chemistry transport model of Chin et al. (1999): (*a*) SO_2 and (*b*) SO_4^{2-}.

to the neighboring oceans and to the free troposphere, while the rest is removed by dry and wet depositions within the source region itself. These authors conclude that dry deposition takes up nearly one third of surface SO_2 emissions directly in the polluted region itself. Thus, although global anthropogenic emissions of SO_2 account for about 70 to 80% of the total emission of sulfate precursors, their contribution to the total sulfate burden is likely to be substantially less.

Yet another interesting result from the model studies is their assessment of the importance of volcanic emissions to the global sulfate burden in the troposphere. Chin and Jacob (1996) have found that this source is a significant contributor to the

TABLE 3 Ranges of Sources, Sinks, Total Mass and Lifetimes of SO_2 and Sulfate from Seven Global Sulfur Models

	Ranges	Median[a]
SO_2		
Total source (Tg S/yr)		95.7
Anthropogenic emission	63.7–92.0	66.5
Biomass burning	2.2–2.9	2.3
Volcanoes	3.4–8.5	5.5
Photochemical production	10.0–24.7	16.9
Total sink (Tg S/yr)		91.2
Gas-phase oxidation	6.1–16.8	9.2
In-cloud oxidation	23.3–55.5	42
Dry deposition	16.0–55.0	35.5
Wet deposition	0–19.9	9.0
Total atmospheric burden (Tg S)	0.2–0.6	0.4
Lifetime (days)	0.6–2.6	1.5
Sulfate		
Total source (Tg S/yr)		50.6
Anthropogenic emission	0–3.5	1.4
Gas-phase production	6.1–16.8	9.2
In-cloud processing	23.3–57.8	40.0
Total sink (Tg S/yr)		50.2
Dry deposition	3.7–17.0	6.7
Wet deposition	34.6–61.0	44.5
Total atmospheric burden (Tg S)	0.3–0.96	0.63
Lifetime (days)	3.9–5.8	4.6

Note: Models are from Langner and Rodhe (1991), Pham et al. (1995), Feichter et al. (1995), Chin et al. (1996), Chuang et al. (1997), Koch et al. (1999), and Barth et al. (1999).

[a] As a result of using median values derived from several different models to define the total source and sink for SO_2 and sulfate, these values do not necessarily balance.

sulfate budget in the middle and upper troposphere. This is due to volcanoes providing direct injection of sulfur gases to upper altitudes where species such as SO_2 typically have lifetimes an order of magnitude longer than that in the boundary layer. The major impact of global volcanic emissions at high altitude is a conclusion also reached by Graf et al. (1998). These investigators found that the global mean radiative forcing by volcanic sulfate aerosols was actually comparable to anthropogenic aerosols.

Table 3 summarizes the global SO_2-sulfate budget results based on the modeling results from several groups (e.g. Langner and Rodhe, 1991; Pham et al., 1995; Feichter et al., 1996; Chin et al., 1996, 2000; Chuang et al., 1997; Koch et al., 1999; Barth et al., 2000). As one might expect, there are a number of differences

among these models, especially in their handling of meteorological fields and parameterizations. Even so, all still agree on certain key points. For example, all models assign 70 to 75% of sulfate precursor emissions to anthropogenic activities; 30 to 45% of the primary SO_2 is also estimated to be removed by dry deposition; and finally, it is agreed that concerning the oxidation of SO_2 to sulfate, 65 to 85% of the total is dominated by in-cloud processes. Among the important areas where significant uncertainties still exist is that of fully understanding the levels of SO_2 and sulfate in remote marine regions. As discussed in Section 3, of particular concern is assessing the relative contributions at free tropospheric altitudes of sulfate derived from DMS oxidation versus that from volcanoes and long-range transport of surface-generated continental sources.

5 STRATOSPHERIC SULFUR

The presences of sulfur in the stratosphere in the form of a sulfate aerosol layer, or Junge layer, was first reported in the early 1960s (Junge et al., 1961). Since its discovery, there have been substantial advances in understanding the effects of stratospheric sulfur on climate and atmospheric chemistry. The primary importance of stratospheric sulfur is that it affects Earth's radiative balance. Aerosols can directly scatter incoming solar radiation back to space. This results in a cooling of Earth's surface. By absorbing outgoing infrared radiation, however, they can also cause a warming of the stratosphere. These effects have been observed after major volcanic eruptions (e.g., Labutzke and McCormick, 1992). Stratospheric aerosols can also have an indirect effect on the radiative balance by acting as CCN. For example, they are involved in forming polar stratospheric clouds (PSCs) and possibly in the development of large-scale cirrus clouds. In addition, stratospheric aerosols may play a significant role in stratospheric chemistry by providing surfaces upon which heterogeneous reactions take place. Such reactions appear to be centrally important as a means of modulating stratospheric ozone levels (see, e.g., Hofmann and Solomon, 1989).

The composition of stratospheric aerosols appears to be mainly sulfate (Rosen, 1971). The most likely source of this sulfate is oxidation of SO_2 to form $H_2SO_4(g)$ as discussed above in Section 3. This oxidation step would then be followed by nucleation and condensation. On the basis of numerous observations of stratospheric aerosols over the past 30 years, volcanic eruptions that inject large amounts of SO_2 directly into the stratosphere are now believed to be one of the dominant sources of stratospheric sulfate aerosols. However, because of the presence of a persistent background of aerosol even during periods when no major volcanic eruptions occurred, there has been considerable speculation concerning other possible sources of this aerosol.

The importance of carbonyl sulfide (OCS) as a stratospheric aerosol source was first proposed by Crutzen (1979). As noted in Section 2, carbonyl sulfide is the most abundant sulfur compound in the atmosphere. It is emitted at Earth's surface by natural and anthropogenic sources, and it is also formed by the oxidation of carbon

disulfide (CS_2) (Chin and Davis, 1993). Recall, however, that because of its chemical inertness in the troposphere, it is found to have a near uniform mixing ratio (i.e., 500 pptv) throughout this region. Because of this, significant quantities of OCS are transported to the stratosphere where it undergoes photodecomposition and/or oxidation via reactions with $O(^3P)$ atoms and OH radicals. The resulting product SO_2, like that from volcanic injections, is then converted to sulfate aerosol. Early modeling studies supported Crutzen's hypothesis and showed that the flux of OCS into the stratosphere was sufficient to maintain the background sulfate aerosol layer.

Twenty years later, with a far more extensive set of OCS atmospheric observations and with improved laboratory reaction rate data, Chin and Davis (1995) reanalyzed the stratospheric significance of OCS as a source of background sulfate aerosols. They compared the flux of OCS calculated in a one-dimensional model with the flux needed to sustain the background aerosol level. Historically, the background level has been estimated from the ratio of background aerosol mass to aerosol lifetime. Departing from earlier analyses, Chin and Davis (1995) found that OCS could provide only 20 to 50% of the required sulfur. This conclusion was based on two important insights: (1) The so-called background aerosol layer observed during volcanic quiescent periods still contained a significant amount of residual volcanic aerosol, and (2) important sources other than OCS quite likely were also contributing to background sulfate aerosol levels. Although a more recent one-dimensional model study, which included microphysical processes, proposed that a sustainable background sulfate layer could indeed be maintained by OCS oxidation (Zhao et al., 1995), Weisenstein et al. (1997) report results that are much closer to those given earlier by Chin and Davis. Weisenstein et al., using a global two-dimensional model, found that OCS oxidation could only account for half of the background sulfur loading. They also found that convective transport of SO_2 in the tropical troposphere could provide the other half of the background sulfate aerosol. In a still more recent study Mills et al. (1999), based on new measurements by Wilson et al. (1999), have suggested that in addition to SO_2, tropospheric sulfate aerosol at the tropopause could make a significant contribution to the sulfate loading of the stratosphere during quiescent periods. As shown in Figures 9*a* and 9*b*, the concentrations of SO_2 and sulfate at the Northern Hemisphere tropopause can reach 50 and 100 pptv, respectively. Quite clearly, there are still important aspects of the so-called background aerosol layer issue that are still unresolved.

The variability in background sulfate aerosol levels has also drawn considerable attention since human activities may have already perturbed the natural background level. For example, there have been reports published indicating that we could be experiencing as much as a 6 to 8% per year increase in background levels. Although the initial speculation was focused on these increases being tied to anthropogenic emissions of OCS, upon further reflection this explanation has been largely rejected. This follows from the fact that there has been no significant long-term trend in OCS concentrations in the troposphere over the last 20 years. Hofmann (1991) has noted, however, that the increase in background aerosol mass is closely related to increases in sulfur emissions from high-altitude aircraft. Another possible anthropogenic

source would involve the direct transport of SO_2 and sulfate from the troposphere, as discussed earlier. The anthropogenic fraction of sulfate in the Northern Hemisphere's upper troposphere can vary from 20% in January to 60 to 80% in July (Chin et al., 2000). Thus, an increase in anthropogenic SO_2 emissions could have made an impact on the stratospheric aerosol level.

A quite different perspective on background stratospheric aerosol trends has been put forward by Chin and Davis (1995). They have raised the question whether one can even reliably define a baseline value for aerosol in an environment that is continually being disturbed by new volcanic injections of sulfur. They point out that there were only 2 years in the 10-year record cited by Sedlacek et al. (1983) and 2 years in the 18-year observations by Hofmann (1990) that could be identified as "volcanic quiescent" periods. In neither case, however, was it possible to convincingly show that the aerosol or sulfate levels observed during these periods were free of any significant volcanic influence. Given the multiyear residence time of volcanic aerosols and the frequency of minor volcanic injections, Chin and Davis (1995) argued that overall there still remains a serious question whether a true background sulfate aerosol level (i.e., one largely uninfluenced by volcanic emissions) has as yet been observed. Thus, critical to any future analyses designed to show the role of tropospheric sulfur compounds in forming stratospheric sulfur aerosol will be the further elucidation of the volcanic component of the so-called background aerosol layer.

REFERENCES

Andreae, M. O., and W. A. Jaeschke, Exchange of sulphur between biosphere and atmosphere over temperate and tropical regions, in R. W. Howarth, J. W. B. Stewart, and M. V. Ivanov (Eds.), *Sulphur Cycling on the Continents: Wetlands, Terrestrial Ecosystems, and Associated Water Bodies*, SCOPE 48, Wiley, Chichester, 1992, pp. 27–61.

Bandy, A. R., D. C. Thomton, B. W. Blomquist, S. Chen, T. P. Wade, J. C. Ianni, G. M. Mitchell, and W. Nadler, Chemistry of dimethyl sulfide in the equatorial Pacific atmosphere, *Geophys. Res. Lett.*, *23*, 741–744, 1996.

Bandy, A. R., D. C. Thomton, and A. R. Driedger III, Airbome measurements of sulfur dioxide, dimethyl sulfide, carbon disulfide, and carbonyl sulfide by isotope dilution gas chromatography/mass spectrometry, *J. Geophys. Res.*, *98*, 23423–23433, 1993.

Barrie, L. A., M. P. Olson, and K. K. Oikawa, The flux of anthropogenic sulphur into the Arctic from mid-latitudes in 1979/80, *Atmos. Environ.*, *18*, 2711–2722, 1989.

Barth, M. C., P. J. Rasch, J. T. Kiehl, C. M. Benkovitz, and S. E. Schwartz, Sulfur chemistry in the NCAR CCM: Description, evaluation, features and sensitivity to aqueous chemistry, *J. Geophys. Res.*, *105*, 1387–1415, 2000.

Bates, T. S., J. A. Calhoun, and P. K. Quinn, Variations in the methanesulfonate to sulfate molar ratio in submicrometer marine aerosol particles over the South Pacific Ocean, *J. Geophys. Res.*, *97*, 9859–9865, 1992a.

Bates, T. S., B. K. Lamb, A. Guenther, J. Dignon, and R. E. Stoiber, Sulfur emissions to the atmosphere from natural sources, *J. Atmos. Chem.*, *14*, 315–337, 1992b.

Bates, T. S., B. J. Huebert, J. L. Gras, F. B. Griffiths, and P. A. Durkee, The international Global Atmospheric Chemistry (IGAC) Project's First Aerosol Characterization Experiment (ACE 1): Overview, *J. Geophys. Res.*, *103*, 16297–16318, 1998.

Berresheim, H., Biogenic sulfur emissions from the Subantarctic and Antarctic oceans, *J. Geophys. Res.*, *92*, 13245–13262, 1987.

Berresheim, H., and F. L. Eisele, Sulfur chemistry in the Antarctic Troposphere Experiment: An overview of project SCATE, *J. Geophys. Res.*, *103*, 1619–1627, 1998.

Berresheim, H., F. L. Eisele, D. J. Tanner, D. S. Covert, L. McInnes, and D. C. Ramsey-Bell, Atmospheric sulfur chemistry and cloud condensation nuclei (CCN) concentrations over the northeastern Pacific coast, *J. Geophys. Res.*, *98*, 12701–112711, 1993.

Berresheim, H., P. Wine, and D. Davis, Sulfur in the atmosphere, in H. B. Singh (Ed.), *Composition, Chemistry, and Climate of the Atmosphere*, Van Nostrand Reinhold, New York, 1995, pp. 251–307.

Chameides, W. L., and D. D. Davis, The free radical chemistry of cloud droplets and its impact upon the composition of rain, *J. Geophys. Res.*, *87*, 4863–4877, 1982.

Chameides, W. L., and A. W. Stelson, Aqueous-phase chemical processes in deliquescent sea-salt aerosols: A mechanism that couples the atmospheric cycles of S and sea salt, *J. Geophys. Res.*, *97*, 20565–20580, 1992.

Charlson, R. J., S. E. Schwartz, J. M. Hales, R. D. Cess, J. A. Coakley, J. E. Hansen, and D. J. Hofmann, Climate forcing by anthropogenic aerosols, *Science*, *255*, 423–430, 1992.

Chen, G., D. D. Davis, P. Kasibhatla, A. R. Bandy, D. C. Thornton, B. J. Huebert, A. D. Clarke, and B. Blomquist, A study of DMS oxidation in the tropics: Comparison of Christmas Island field observations of DMS, SO_2, and DMSO with model simulations, *J. Atmos. Chem.*, *37*, 137–160, 2000.

Chin, M., and D. D. Davis, Global sources and sinks of OCS and CS_2 and their distributions, *Global Biogeochem. Cycles*, *7*, 321–337, 1993.

Chin, M., and D. D. Davis, A reanalysis of carbonyl sulfide as a source of stratospheric background sulfur aerosol, *J. Geophys. Res.*, *100*, 8993–9005, 1995.

Chin, M., and D. J. Jacob, Anthropogenic and natural contributions to tropospheric sulfate: A global model analysis, *J. Geophys. Res.*, *101*, 18691–18699, 1996.

Chin, M., D. J. Jacob, G. M. Gardner, M. S. Foreman-Fowler, P. A. Spiro, and D. L. Savoie, A global three-dimensional model of tropospheric sulfate, *J. Geophys. Res.*, *101*, 18667–18690, 1996.

Chin, M., R. Rood, S.-J. Lin, J. F. Muller, and A. Thompson, Atmospheric sulfur cycle simulated in the global model GOCART: Model description and global properties, *J. Geophys. Res.*, *105*, 24671–24687, 2000.

Chuang, C. C., J. E. Penner, K. E. Taylor, A. S. Grossman, and J. J. Walton, An assessment of the radiative effects of anthropogenic sulfate, *J. Geophys. Res.*, *102*, 3761–3778, 1997.

Clarke, A. D., Atmospheric nuclei in the Pacific midtroposphere: Their nature, concentration and evolution, *J. Geophys. Res.*, *98*, 20633–20647, 1993.

Corbett, J. J., and P. S. Fischbeck, Emissions from ship, *Science*, *278*, 823–824, 1997.

Corbett, J. J., P. S. Fischbeck, and S. N. Pandis, Global nitrogen and sulfur inventories for oceangoing ships, *J. Geophys. Res.*, *104*, 457–3470, 1999.

Crutzen, P. J., The possible importance of OCS for the sulfate layer of the stratosphere, *Geophys. Res. Lett.*, *3*, 73–76, 1979.

Davis, D. D., G. Chen, A. Bandy, D. Thornton, F. Eisele, L. Mauldin, D. Tanner, D. Lenschow, H. Fuelberg, B. Huebert, J. Heath, A. Clarke, and D. Blake, Dimethyl sulfide oxidation in the equatorial Pacific: Comparison of model simulations with field observations for DMS, SO_2, H_2SO_4(g), MSA(g), MS, and NSS, *J. Geophys. Res.*, *104*, 5765–5784, 1999.

Davis, D. D., (unpublished results) in text.

Davis, D. D., G. Chen, P. Kasibhatla, A. Jefferson, D. Tanner, F. Eisele, D. Lenschow, W. Neff, and H. Berresheim, DMS oxidation in the Antarctic marine boundary layer I: Comparison of model simulations and field observations for DMS, DMSO, $DMSO_2$, H_2SO_4(g), MSA(g), and MSA(p), *J. Geophys. Res.*, *103*, 1657–1678, 1998.

De Bruyn, W. J., T. S. Bates, J. M. Cainey, and E. S. Saltzman, Shipboard measurements of dimethyl sulfide and SO_2 southwest of Tasmania during the First Aerosol Characterization Experiment (ACE 1), *J. Geophys. Res.*, *103*, 16703–16711, 1998.

Feichter, J., E. Kjellstrom, H. Rodhe, F. Dentener, J. Lelieveld, and G.-J. Roelofs, Simulation of the tropospheric sulfur cycle in a global climate model, *Atmos. Environ.*, *30*, 1693–1708, 1996.

Finlayson-Pitts, B., and J. N. Pitts, *Atmospheric Chemistry: Fundamentals and Experimental Techniques*, Wiley, New York, 1986.

Graf, H.-F., B. Langmann, and J. Feichter, The contribution of Earth degassing to the atmospheric sulfur budget, *Chem. Geol.*, *147*, 131–145, 1998.

Grosjean, D., Photooxidation of methyl sulfide, ethyl sulfide, and methanethiol, *Environ. Sci. Technol.*, *18*, 460–468, 1984.

Hameed, S., and J. Dignon, Global emissions of nitrogen and sulfur oxides in fossil fuel combustion: 1970–1986, *J. Air Waste Mgmt. Assoc.*, *42*, 159–163, 1992.

Hatakeyama, S., K. Izumi, and H. Akimoto, Yield of SO_2 and formation of aerosol in the photo-oxidation of DMS under atmospheric conditions, *Atmos. Environ.*, *19*, 135–141, 1985.

Hoell, Jr., J. M., D. D. Davis, D. J. Jacob, M. O. Rogers, R. B. Newell, H. E. Fuelberg, R. J. McNeal, J. L. Raper, and R. J. Bendura, Pacific Exploratory Mission in the tropical Pacific: PEM—Tropics A, August–September, 1996, *J. Geophys. Res.*, *104*, 5567–5583, 1999.

Hofmann, D. J., Increase in the stratospheric background sulfuric acid aerosol mass in the past 10 year, *Science*, *248*, 996–1000, 1990.

Hofmann, D. J., Aircraft sulphur emissions, *Nature*, *349*, 659, 1991.

Hofmann, D. J., and S. Solomon, Ozone destruction through heterogeneous chemistry following the eruption of El Chichon, *J. Geophys. Res.*, *94*, 5029–5041, 1989.

Hynes, A. J., and P. H. Wine, The atmospheric chemistry of dimethylsulfoxide (DMSO) kinetics and mechanism of the OH + DMSO reaction, *J. Atmos. Chem.*, *24*, 23–37, 1996.

Hynes, A. J., P. H. Wine, and D. H. Semmes, Kinetics and mechanism of OH reactions with organic sulfides, *J. Phys. Chem.*, *90*, 4148–4156, 1986.

Jefferson, A., D. J. Tanner, F. L. Eisele, D. D. Davis, G. Chen, J. Crawford, J. W. Huey, A. L. Torres, and H. Berresheim, OH photochemistry and methane sulfonic acid formation in the coastal Antarctic boundary layer, *J. Geophys. Res.*, *103*, 1647–1656, 1998.

Junge, C. E., C. W. Chagnon, and J. E. Manson, Stratospheric aerosols, *J. Meteorol.*, *18*, 81–108, 1961.

Koch, D., D. Jacob, I. Tegen, D. Rind, and M. Chin, Tropospheric sulfur simulation and sulfate direct radiative forcing in the Goddard Institute for Space Studies general circulation model, *J. Geophys. Res.*, *104*, 23799–23822, 1999.

Kreidenweis, S. M., and J. H. Seinfeld, Nucleation of sulfuric acid–water and methanesulfonic acid–water solution particles: Implications for the atmospheric chemistry of organosulfur species, *Atmos. Environ.*, *22*, 283–296, 1988.

Krouse, H. R., and R. G. L. McCready, Reductive reactions in the sulfur cycle, in P. A. Trudinger and D. J. Swaine (Eds.), *Biogeochemical Cycling of Mineral-Forming Elements*, Elsevier, Amsterdam, 1979, pp. 315–368.

Labutzke, K., and M. P. McCormick, Stratospheric temperature increase due to Pinatobo aerosols, *Geophys. Res. Lett.*, *19*, 207–210, 1992.

Langner, J., and H. Rodhe, A global three-dimensional model of the tropospheric sulfur cycle, *J. Atmos. Chem.*, *13*, 225–263, 1991.

Lee, Y.-N., and X. Zhou, Aqueous reaction kinetics of ozone and dimethlysulfide and its atmospheric implications, *J. Geophys. Res.*, *99*, 3597–3605, 1994.

Lelieveld, J., and J. Heintzenberg, Sulfate cooling effect on climate through in-cloud oxidation of anthropogenic SO_2, *Science*, *258*, 117–120, 1992.

Lovejoy, E. R., D. R. Hanson, and L. G. Huey, Kinetics and products of the gas-phase reaction of SO_3 with water, *J. Phys. Chem.*, *100*, 19911–19916, 1996.

Martin, L. R., Kinetic studies of sulfite oxidation in aqueous solution, in J. G. Calvert (Ed.), *SO_2, NO and NO_2 Oxidation Mechanisms: Atmospheric Considerations*, Butterworth, Boston, 1984, pp. 63–100.

Mauldin III, R. L., D. J. Tanner, and F. L. Eisele, Measurements of OH during PEM—Tropics A, *J. Geophys. Res.*, *104*, 5817–5827, 1999a.

Mauldin III, R. L., D. J. Tanner, J. A. Heath, B. J. Huebert, and F. L. Eisele, Observations of H_2SO_4 and MSA during PEM—Tropics, *J. Geophys. Res.*, *104*, 5801–5816, 1999b.

Mills, J. M., O. B. Toon, and S. Solomon, A microphysical analysis of non-volcanic sources of atmospheric sulfate, in *EOS Trans.*, AGU, 1999 San Francisco Fall Meeting, Vol. 80, 1999, p. F169.

Niki, H., P. D. Maker, C. M. Savage, and L. P. Breitenbach, An FTIR study of the mechanism for the reaction HO + CH_3SCH_3, *Int. J. Chem. Kinet.*, *15*, 647–654, 1983.

Nowak, J., D. D. Davis, G. Chen, F. Eisek, D. Tanner, L. Mauldin III, C. Cantrell, E. Koscinch, A. Baudy, D. Thornton, and A. Clarke, *Geophys. Res. Ltt.*, p2201–2204, 2001.

Pham, M., J.-F. Müller, G. P. Brasseur, C. Granier, and G. Mégie, A three-dimensional study of the tropospheric sulfur cycle, *J. Geophys. Res.*, *100*, 26061–26092, 1995.

Rosen, J. M., The boiling point of stratospheric aerosols, *J. Appl. Meteor.*, *10*, 1044–1046, 1971.

Saltzman, E. S., D. L. Saugie, R. G. Zika, and J. M. Prosporo, Methane sulfonic acid and non-sea-salt sulfate in Pacific air: regional and seasonal variations. *J. Atmos. Chem.*, 4, 227–240, 1986.

Saltzman, E. S., S. A. Yvon, and P. A. Matrai, Low-level detection of atmospheric sulfur dioxide measurement using HPLC/fluorescence detection, *J. Atmos. Chem.*, *17*, 73–90, 1993.

Savoie, D. L., and J. M. Prospero, Comparison of oceanic and continental sources of non-sea-salt sulphate over the Pacific Ocean, *Nature*, *339*, 685–687, 1989.

Savoie, D. L., J. M. Prospero, and E. S. Saltzman, Nitrate, non sea-salt sulfate and nitrate in trade wind aerosols at Barbados: Evidence for long range transport, *J. Geophys. Res.*, *94*, 5069–5080, 1989.

Sedlacek, W. A., E. J. Mroz, A. L. Lazrus, and B. W. Gandrud, A decade of stratospheric sulfate measurements compared with observations of volcanic eruptions, *J. Geophys. Res.*, *88*, 3741–3776, 1983.

Sievering, H., J. Boatman, E. Gorman, Y. Kim, L. Anderson, G. Ennis, M. Luria, and S. Pandis, Removal of sulphur from the marine boundary layer by ozone oxidation in sea-salt aerosol, *Nature*, *360*, 571–573 , 1992.

Singh, H. B., A. Thakur, Y. E. Chen, and M. Kanakidou, Tetrachloroethene as an indicator of low Cl atom concentrations in the troposphere, *Geophys. Res. Lett.*, *23*, 1529–1532, 1996.

Spiro, P. A., D. J. Jacob, and J. A. Logan, Global inventory of sulfur inventory of sulfur emissions with $1° \times 1°$ resolution, *J. Geophys. Res.*, *97*, 6023–6036, 1992.

Thornton, D. C., A. R. Bandy, B. W. Bloomquist, A. R. Driedger, and T. P. Wade, Sulfur dioxide distribution over the Pacific Ocean 1991–1996, *J. Geophys. Res.*, *104*, 5845–5854, 1999.

Turnipseed, A. A., and A. R. Ravishankara, The atmospheric oxidation of dimethyl sulfide: Elementary steps in a complex mechanism, in G. Restelli, and G. Angeletti (Eds.), *Dimethylsulphide: Oceans, Atmosphere and Climate*, Kluwer, Dordrecht, 1993, pp. 185–196.

Urbanski, S. P., R. E. Stickel, and P. H. Wine, Mechanistic and kinetic study of the gas-phase reaction of hydroxyl radical with dimethyl sulfoxide, *J. Phys. Chem.*, *102*, 10522–10529, 1998.

Weisenstein, D. K., G. K. Yue, M. K. W. Ko, N. D. Sze, J. M. Rodriguez, and C. J. Scott, A two-dimensional model of sulfur species and aerosols, *J. Geophys. Res.*, *102*, 13019–13035, 1997.

Wilson, J. O., O. A. Brock, and J. J. Jonsson, In situ measurements of aerosol properties in the upper troposphere and lower stratosphere: Is the Pinatubo aerosol still decaying? in *EOS. Trans.*, AGU, 1999 Fall Meeting, Vol. 80, 1999, p. F169.

Yin, F., D. Grossjean, and J. H. Seinfeld, Photooxidation of dimethyl sulfide and dimethyl disulfide, *J. Atmos. Chem.*, *11*, 309–399, 1990.

Yvon, S. A., E. S. Saltzman, D. J. Cooper, T. S. Bates, and A. M. Thompson, Atmospheric sulfur cycling in the tropical Pacific marine boundary layer (12°S, 135°W): A comparison of field data and model results 2. Sulfur dioxide, *J. Geophys. Res.*, *101*, 6911–6918, 1996.

Zhao, J.-X, R. P. Turco, and O. B. Toon, A model simulation of volcanic aerosol evolution in the stratosphere, *J. Geophys. Res.*, *100*, 7315–7328, 1995.

CHAPTER 9

CONVECTIVE TRANSPORT

KENNETH E. PICKERING

1 INTRODUCTION

In the early 1980s it was recognized that observed free tropospheric mixing ratios of some trace gases could not be explained simply by large-scale transport and eddy diffusion. Crutzen and Gidel (1983), Gidel (1983), and Chatfield and Crutzen (1984) hypothesized that convective clouds played an important role in rapid atmospheric vertical transport of trace species and tested parameterizations of convective transport in atmospheric chemical models. At nearly the same time evidence was shown of venting of the boundary layer by shallow fair weather cumulus clouds (e.g., Greenhut et al., 1984; Greenhut, 1986). Field experiments were conducted in 1985 that resulted in verification of the hypothesis that deep convective clouds are instrumental in atmospheric transport of trace constituents (Dickerson et al., 1987; Garstang et al., 1988). Once pollutants are lofted to the middle and upper troposphere, they typically have a much longer chemical lifetime and with the generally stronger winds at these altitudes they can be transported large distances from their source regions. Photochemical reactions occur during this long-range transport. Pickering et al. (1990) demonstrated that venting of boundary layer pollutants by convective clouds (both shallow and deep) causes enhanced ozone production in the free troposphere. Therefore, convection aids in the transformation of local pollution into a contribution to global atmospheric pollution.

Field studies have established that downward transport of larger O_3 and NO_x mixing ratios from the free troposphere to the boundary layer is an important process over the remote oceans (e.g., Piotrowicz et al., 1991), as well as the upward transport of very low O_3 mixing ratios from the boundary layer to the upper troposphere (Kley et al., 1996). Global modeling by Lelieveld and Crutzen (1994) suggests that the downward mixing of O_3 into the boundary layer is the dominant global effect of

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

deep convection. Some indications of downward transport of O_3 from higher altitudes (possibly from the stratosphere) in the anvils of thunderstorms have been observed (Dickerson et al., 1987; Poulida et al., 1996; Suhre et al., 1997). Ozone is most effective as a greenhouse gas in the vicinity of the tropopause. Therefore, changes in the vertical profile of O_3 in the upper troposphere caused by deep convection have important radiative forcing implications for climate.

More detailed discussion of observations of convective transport are presented in Section 2. Simulation of convective transport in cloud-resolving models and its parameterization in larger-scale models is discussed in Section 3, as well as implications for O_3 production following convective redistribution.

2 OBSERVATIONS

Venting by Nonprecipitating Cumulus Clouds

Some fraction of shallow fair weather cumulus clouds actively vent boundary layer pollutants to the free troposphere (Stull, 1985). The first airborne observations of this phenomenon were conducted by Greenhut et al. (1984) over a heavily urbanized area, measuring the in-cloud flux of ozone in a relatively large cumulus cloud. An extension of this work was reported by Greenhut (1986) in which data from over 100 aircraft penetrations of isolated nonprecipitating cumulus clouds over rural and suburban areas were obtained. Ching and Alkezweeny (1986) reported tracer (SF_6) studies associated with nonprecipitating cumulus (fair weather cumulus and cumulus congestus). Their experiments showed that the active cumulus clouds transported mixed layer air upward into the overlying free troposphere and suggested that active cumuli can also induce rapid downward transport from the free troposphere into the mixed layer. A UV-DIAL (ultraviolet differential absorption lidar) provided space-height cross sections of aerosols and ozone over North Carolina in a study of cumulus venting reported by Ching et al. (1988). Data collected on evening flights showed regions of cloud debris containing aerosol and ozone in the lower free troposphere in excess of background, suggesting that significant vertical exchange had taken place during afternoon cumulus cloud activity. Efforts have also been made to estimate the vertical transport by ensembles of nonprecipitating cumuli in regional chemical transport models (e.g., Vukovich and Ching, 1990).

Deep Convection

Midlatitudes. The first unequivocal observations of deep convective transport of boundary layer pollutants to the upper troposphere were documented by Dickerson et al. (1987). Instrumentation aboard three research aircraft measured CO, O_3, NO, NO_x, NO_y, and hydrocarbons in the vicinity of an active mesoscale convective system near the Oklahoma/Arkansas border during the 1985 PRE-STORM experiment. Anvil penetrations about 2 h after maturity found greatly enhanced mixing

ratios of all of the aforementioned species compared with outside of the cloud. Among the species measured, CO is the best tracer of upward convective transport because it is produced primarily in the boundary layer and has an atmospheric lifetime much longer than the time scale of a thunderstorm. In the observed storm CO measurements exceeded 160 ppbv as high as 11 km, compared with ~ 70 ppbv outside of the cloud (Fig. 1*a*). Nonmethane hydrocarbons (NMHC) with moderate lifetimes can also trace convective transport from the boundary layer. Ozone can also be an indicator of convective transport; in the polluted troposphere large ozone values will indicate upward transport from the boundary layer, but in the clean atmosphere such values are indicative of downward transport from the uppermost troposphere or lowermost stratosphere. In this case measured ozone in the upper rear portion of the anvil peaked at 98 ppbv, while boundary layer values were only ~ 65 ppbv (Fig. 1*b*). It is likely that some higher ozone stratospheric air mixed into the anvil. Because lightning makes major contributions to reactive nitrogen in thunderstorms, NO_x measurements are unsuitable as a convective tracer.

The large amount of vertical trace gas transport noted by Dickerson et al. (1987) cannot, however, be extrapolated to all convective cells. Pickering et al. (1988) reported airborne measurements of trace gases taken in the vicinity of a line of towering cumulus and cumulonimbus clouds that also occurred during PRE-STORM. In this case trace gas mixing ratios in the tops of these clouds were near ambient levels. Meteorological analyses showed that these clouds were located above a cold front that prevented entry of air from the boundary layer directly below or near the clouds. Instead, the air entering these clouds likely originated in the layer immediately above the boundary layer, which was quite clean. Enhanced values of ozone precursor gases were found in the upper troposphere during another PRE-STORM flight conducted in clear air (Pickering et al., 1989). These observations were identified through correlation analysis as indicative of air with a recent boundary layer source and were traced through back trajectory analysis to deep convection that occurred 600 km upstream. Luke et al. (1992) summarized the air chemistry data from all 18 flights during PRE-STORM by categorizing each case according to synoptic flow patterns. Storms in the maritime flow regime transported large amounts of CO, O_3, and NO_y into the upper troposphere, with the midtroposphere remaining relatively clean. During frontal passages a combination of stratiform and convective clouds mixed pollutants more uniformly into the middle and upper levels; high mixing ratios of CO were found at all altitudes.

Other flights in the vicinity of convective storms over the continental United States were reported by Kleinman and Daum (1991), showing a strong decrease of aerosol particles and water vapor with altitude. However, CO and NO_y were more uniformly distributed in the vertical. Plumelike features, attributed to convective outflow, were noted at high altitude in which mixing ratios of boundary layer pollutants increased by 50% or more above background over a distance of several kilometers. Within these features aerosols and water vapor were enhanced over background values, but these soluble substances were always depleted relative to the insoluble species such as CO, suggesting in-cloud removal of the soluble material.

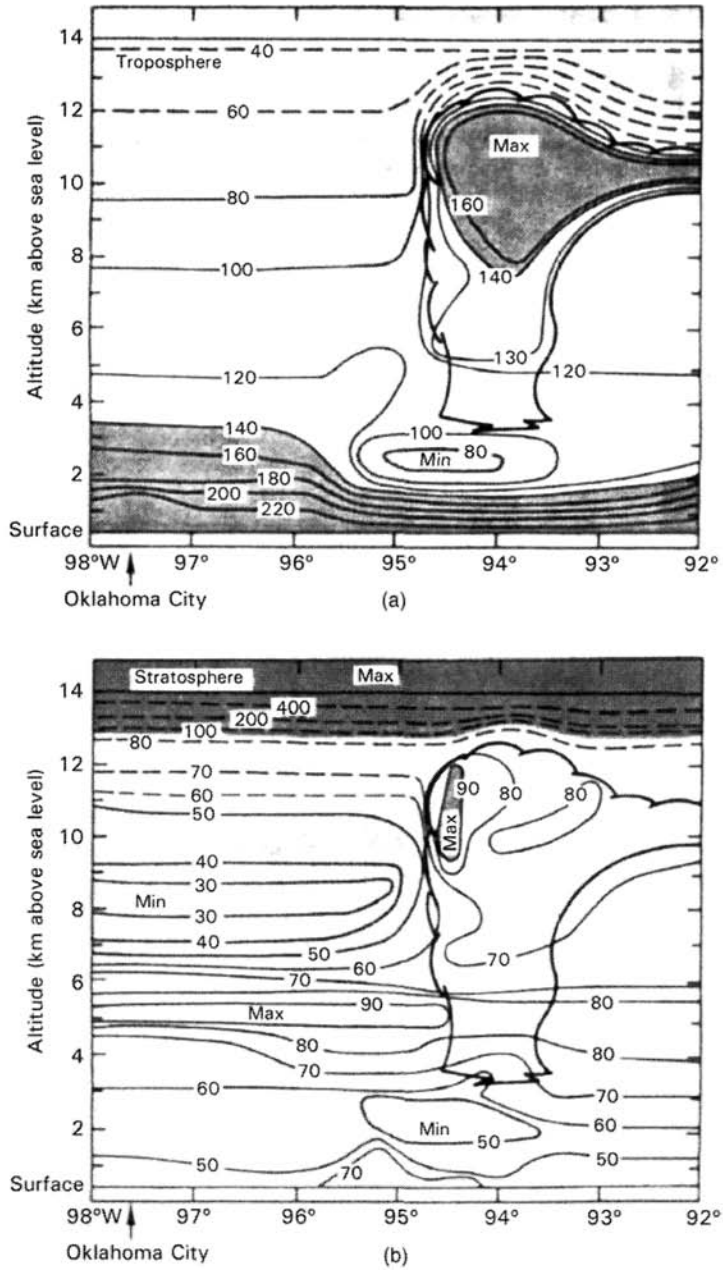


Figure 1 (*a*) Contour plot of CO mixing ratios (ppbv) observed in and near the June 15, 1985, mesoscale convective complex in eastern Oklahoma. Heavy line shows the outline of the cumulonimbus cloud. Dark shading indicates high CO and light shading indicates low CO. Dashed contour lines are plotted according to climatology since no direct measurements were made in that area. (*b*) Same as (*a*) but for ozone (ppbv). From Dickerson et al. (1987).

Poulida et al. (1996) reported observations taken prior to, in, and around a squall line over North Dakota that evolved into a mesoscale convective complex. In this case the anvil extended well into what used to be the stratosphere. Air in the anvil was characterized by low concentrations of O_3 (Fig. 2) and high CO, NO, and NO_y, relative to outside the cloud. This layer of tropospheric air lay above a tongue of stratospheric air, indicating that extensive stratosphere–troposphere exchange had occurred. The flux of O_3 into the troposphere and the fluxes of water vapor, CO, NO_y, and hydrocarbons into the stratosphere were estimated for the storm. If only a small fraction of this material from such anvils remained in the stratosphere, it likely dominates the chemistry of the lower stratosphere in this midlatitude region.

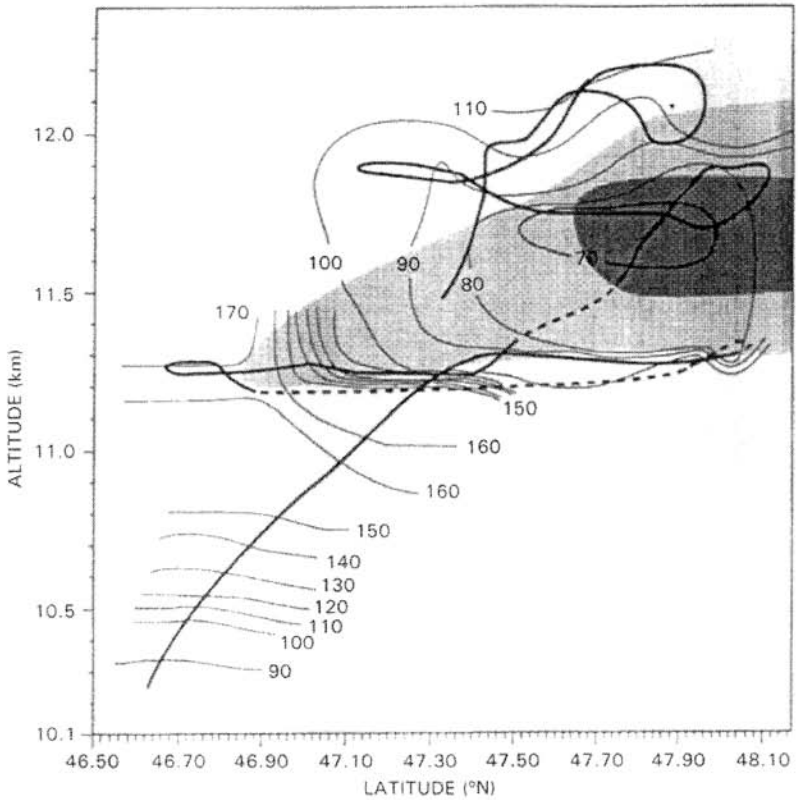


Figure 2 Ozone concentrations in the anvil of a mesoscale convective complex over North Dakota on June 28, 1989. Heavy line indicates flight track projected onto a vertical plane. Thin lines are ozone isopleths every 10 ppbv. Shading shows location of anvil based on aircraft ice particle measurements. Heavier shading indicates greater particle concentrations. From Poulida et al. (1996).

Tropics. Several deep convection experiments with chemical measurements have been conducted in the tropics. Thompson et al. (1997) have summarized many of these results concerning convective transport of trace gases and their consequences for tropospheric ozone production. Garstang et al. (1988) reported measurements taken in front and behind a dry-season squall line over the Amazon rainforest during the NASA ABLE 2A (Amazon Boundary Layer Experiment) project in 1985. The importance of specific processes within the storm (updrafts and downdrafts) as well as the net result of convective transport (atmospheric overturning) were noted. Since the measurements were confined to the lowest 5 km, downward transport of chemical tracers (e.g., ozone) was the most evident feature (Fig. 3). A major emphasis was placed on sampling convective systems during the ABLE 2B wet-season experiment in the same region in 1987. Scala et al. (1990) reported on a locally occurring ABLE 2B convective system, showing that trace gases in the lower troposphere in the wake of the system were well mixed in the vertical. NO measurements behind the storm were greater than ahead of the system, indicating downward transport from above. However, the NO mixing ratios were low enough that ozone production/destruction rates were very small, allowing ozone to be considered a valid tracer of convective

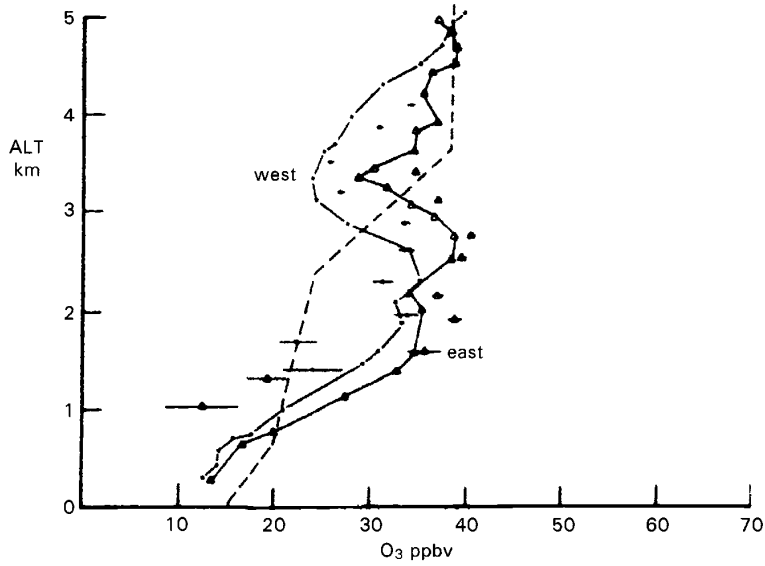


Figure 3 Vertical profiles of ozone concentration along the east and west sides of the August 3, 1985, squall line observed in Brazil in ABLE 2A. The mean profile of ozone from flights in undisturbed weather is shown with a dashed line. Means and standard deviations of ozone from UV-DIAL measurements are shown with symbols and horizontal lines. From Garstang et al. (1988).

transport. Aided by a convective cloud model, Scala et al. (1990) concluded that this system showed that deep undilute convective transport in closed conduits as suggested by Riehl and Simpson (1979) may not necessarily occur in very moist continental tropical systems; the conduits appeared to leak.

Pickering et al. (1996) reported data from a flight of the NASA DC-8 aircraft over Brazil during the TRACE-A (Transport and Atmospheric Chemistry near the Equator—Atlantic) experiment conducted during the biomass burning season of 1992. Outflow from mesoscale convective systems was sampled at 9.5 and 11.3 km showing enhancement of CO mixing ratios typically by a factor of 3 above background (200 to 300 vs. 90 ppbv; see Fig. 4) and significant increases in NO_x and hydrocarbons. Both lightning and transport made important contributions to the enhanced NO_x at cloud outflow levels. Cloud-resolving and regional transport models, a trajectory model, and a photochemical model were used in illustrating the importance of convective events in the ozone budget of the South Atlantic region (see further details in Section 3).

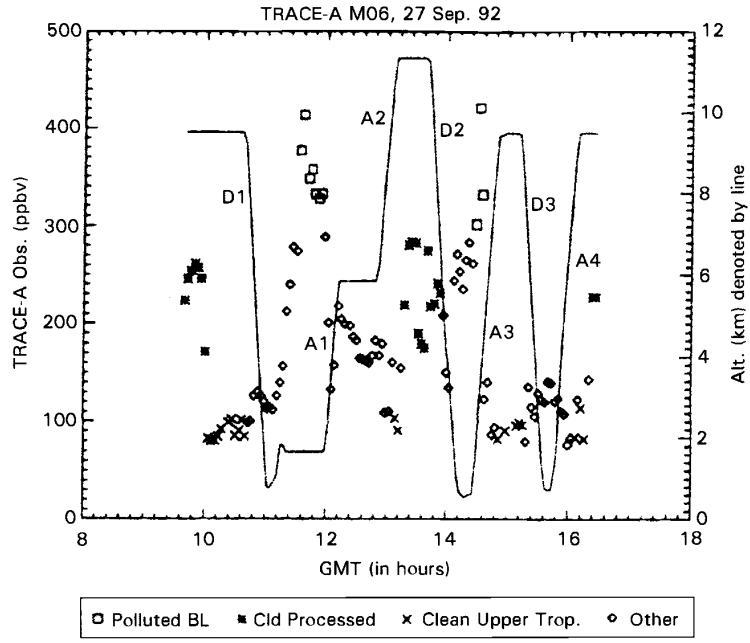


Figure 4 Summary of CO (ppbv) measurements from NASA DC-8 aircraft taken on September 27, 1992, north of Brasilia. Ascents (A) and descents (D) are noted. Three regimes are denoted with indicated symbols and are defined as follows: polluted BL, altitude < 4 km and CO > 300 ppbv; cloud-processed, altitude > 6 km and CO > 150 ppbv; and clean upper troposphere, altitude > 6 km and CO < 120 ppbv. From Pickering et al. (1996).

Over remote marine areas the effects of deep convection on trace gas distributions differ from that over moderately polluted continental regions. Chemical measurements taken by the NASA ER-2 aircraft during the Stratosphere-Troposphere Exchange Project (STEP) off the northern coast of Australia show the influence of very deep convective events. Between 14.5 and 16.5 km on the February 2–3, 1987, flight, perturbations in the chemical profiles were noted that included pronounced maxima in CO, water vapor, CCN and minima of NO_y and ozone (Pickering et al., 1993). Trajectory analysis showed that these air parcels likely were transported from convective cells 800 to 900 km upstream. Very low boundary layer mixing ratios of NO_y and ozone in this remote region were apparently transported upward in the convection. A similar result was noted in CEPEX (Central Equatorial Pacific Experiment; Kley et al., 1996) where a series of ozonesonde ascents showed very low upper tropospheric ozone following deep convection.

Data from convective outflow in the NASA PEM–West A and B experiments (Pacific Exploratory Mission) have been reported by Newell et al. (1996) and by Kawakami et al. (1997). Newell et al. (1996) described sampling of a typhoon in the western Pacific. Boundary layer inflow contained low values of O_3, CO, and hydrocarbons, but high values of dimethylsulfide (DMS). There was no evidence of downward entrainment of stratospheric air into the eye region based on ozone measurements. The DMS data suggested substantial entrainment of boundary layer air into the system, particularly in the eyewall region. Kawakami et al. (1997) reported very low NO_y mixing ratios in the upper troposphere during the February PEM–West B flights between 1°N and 14°N. These measurements were accompanied by very low ozone and large mixing ratios of water vapor and CH_3I, suggesting that the low NO_y values were likely due to convetive transport of tropical marine boundary layer air. Other upper tropospheric measurements showed enhanced NO and high NO_x/NO_y ratios accompanied by low CO, indicative of NO production by lightning.

Danielsen (1993) presented evidence from Darwin, Australia, ER-2 flights in STEP that rapid vertical irreversible transport of lower tropospheric air into the lower tropical statosphere occurs in convective cloud turrets and by large-scale upwelling in tropical cyclones. Suhre et al. (1997) reported O_3 measurements from the tropical Atlantic upper troposphere (10–12 km) taken from commercial aircraft showing mixing ratios of 100 to 500 ppbv at a horizontal scale of 5 to 80 km in the proximity of deep convection. It is hypothesized that there is either direct input of stratospheric O_3 into the anvils of these systems or there is downward convective transport of O_3-rich air that has been transported quasi-isentropically from the extratropical stratosphere.

3 MODELING

Cloud Scale

The Goddard Cumulus Ensemble (GCE) model (Tao and Simpson, 1993) has been used by Pickering et al. (1991, 1992a,b, 1993, 1996), Scala et al. (1990), and

Stenchikov et al. (1996) in the analysis of convective transport of trace gases. The cloud model is nonhydrostatic and contains detailed representation of cloud microphysical processes. Two- and three-dimensional versions of the model have been applied in transport analyses. The initial conditions for the model are usually from a sounding of temperature, water vapor, and winds representative of the region of storm development. Model-generated wind fields can be used to perform air parcel trajectory analyses and tracer advection calculations. Scala et al. (1990) conducted detailed air parcel trajectory analyses for an ABLE 2B storm to investigate flow patterns within the system. In this case the model showed that more than 50% of the air transported to the anvil region originated at or above 6 km, not from the boundary layer via undilute core updrafts. The trajectories also allowed diagnosis of a rotor-type circulation in the low to mid levels of the storm, which was responsible for thorough mixing of the lower troposphere (Fig. 5).

Pickering et al. (1991) used trajectory analyses derived from the GCE model wind fields for the ABLE 2A storm observed by Garstang et al. (1988) to identify air parcels that were undisturbed or modified by the storm (Fig. 6). Tracer transport calculations were performed for CO, O_3, and NO_x, and difference fields showing the changes in mixing ratio of each of these species due to convective transport were computed (Fig. 7). Enhanced values of ozone precursors (NO_x and CO) in a biomass burning haze layer just above the boundary layer were redistributed upward and downward by the storm. Profiles taken from the two-dimensional tracer fields before and after convective transport were used in a one-dimensional photochemical

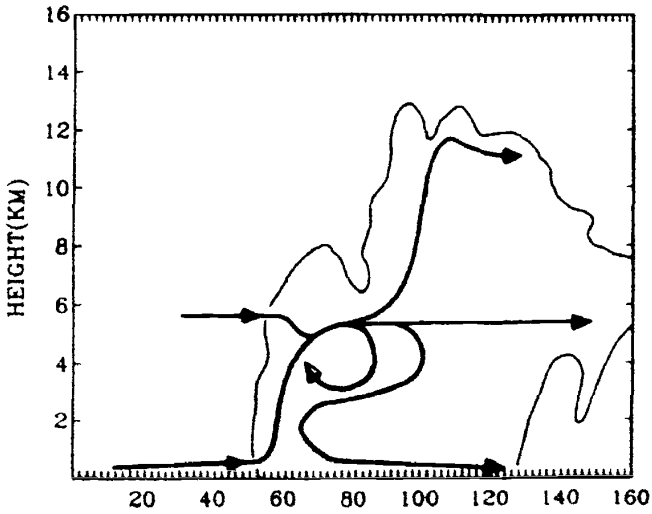


Figure 5 Composite schematic of the predominant transport pathways for the May 6, 1987, ABLE 2B simulated squall convection based on backward and forward trajectory analyses. The model cloud outline at 300 min in the simulation is shown. The horizontal dimension is 80 km. From Scala et al. (1990).

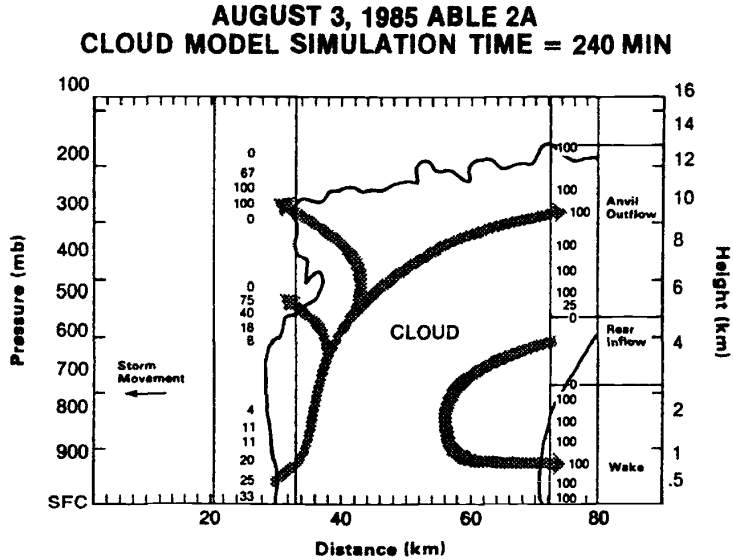


Figure 6 Summary of back trajectories produced by the GCE model for the August 3, 1985, ABLE 2A squall line. Numbers in the vertical column ahead and behind the cloud indicate the percentage of the air at that altitude that is outflow from the cloud. Most of the air pumped out of the boundary layer exits from the anvil (8 to 12 km) and the air in the "wake" has also been processed. Most of the air in the boundary layer ahead of the storm is unperturbed. Arrows indicate main flow paths. From Pickering et al. (1991).

model to estimate ozone production rates. The upward transport of O_3 precursors changed the photochemical tendency of the upper troposphere from that of O_3 destruction to that of production. The same storm dynamics were used in a sensitivity study of convective transport and subsequent free tropospheric O_3 production for conditions of more intense biomass burning pollution (Pickering et al., 1992b). Assuming a pristine middle and upper troposphere prior to convection, enhancements of O_3 production postconvection potentially could be as great as a factor of ~ 50.

Similar methods were used by Pickering et al. (1992a) to examine transport of urban plumes by deep convection. Transport of the Oklahoma City plume by the June 10–11, 1985, PRE-STORM squall line and of the Manaus, Brazil, plume by the April 26, 1987, ABLE 2B squall line were simulated with the two-dimensional GCE model. In the Oklahoma event forward trajectories from the boundary layer at the leading edge of the storm showed that almost 75% of the low-level inflow was transported to altitudes exceeding 8 km. Over 35% of the air parcels reached altitudes over 12 km. For the Amazonian storm, 50% of the trajectories indicated transport to altitudes greater than 12 km. However, nearly 25% of the air parcels indicated air being detrained from the rear of the cloud between 4 and 8 km, and

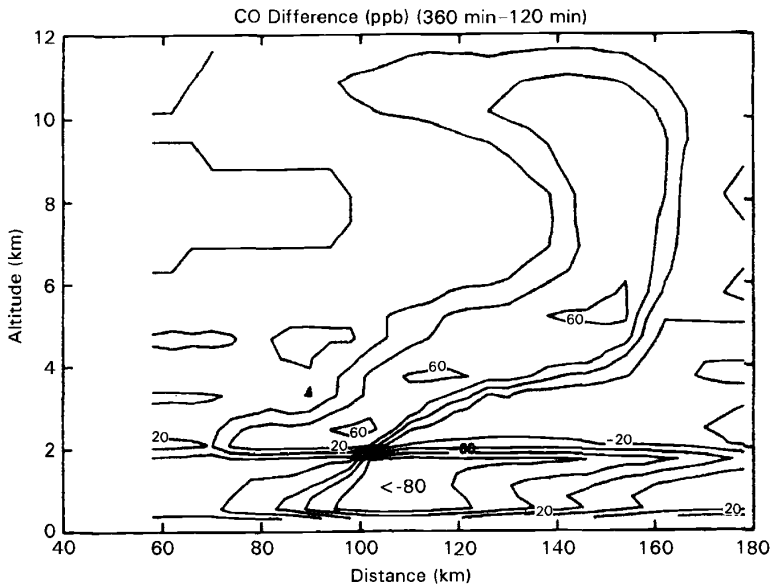


Figure 7 Difference (postconvection minus undisturbed) in model-computed CO tracer concentrations for the August 3, 1995, ABLE 2A squall line. Increases in CO are noted throughout the main updraft region and anvil, and decreases are seen in the downdraft region. From Pickering et al. (1991).

15% became involved in a rotor-type circulation located behind the convective updrafts. In each of these cases tracer transport calculations were performed for CO, NO_x, O_3, and hydrocarbons. The three-dimensional version of the GCE model has also been run for the June 10–11, 1985, PRE-STORM case and for the September 26, 1992, event from TRACE-A. Figure 8 shows the redistributed CO from the rural Oklahoma boundary layer as simulated by the model-generated three-dimensional wind field. Free tropospheric O_3 production enhancement of a factor of 2.5 for Oklahoma rural air and ~ 4 for the Oklahoma City case were calculated, while with a pristine preconvective upper troposphere an enhancement of a factor of 35 was estimated for the Manaus, Brazil, case.

Stenchikov et al. (1996) used the two-dimensional GCE model to simulate the North Dakota storm observed by Poulida et al. (1996). This storm showed the unusual feature of an anvil formed well within the stratosphere. The increase of CO and water vapor above the altitude of the preconvective tropopause was computed in the model. The total mass of CO across the model domain above this level increased by almost a factor of 2 during the convective event. Downward transport of ozone from the stratosphere was noted in the simulation in the rear anvil. Wang et al. (1995) simulated a tropical convective storm observed during CEPEX using the cloud dynamics and cloud transport models of Wang and Chang (1993).

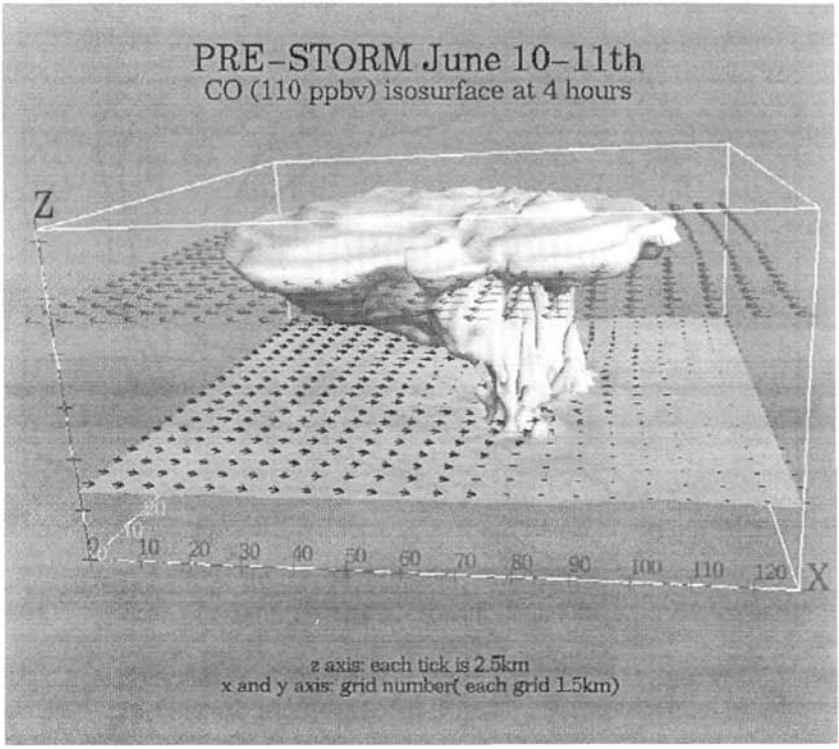


Figure 8 Isosurface of CO mixing ratio (110 ppbv) computed for the June 10, 1985, PRE-STORM squall line over Oklahoma using the three-dimensional GCE model. Measured rural CO mixing ratios used as initial conditions.

The simulated cloud tower extended into the lower stratosphere and a widespread anvil was produced. Intense mixing of boundary layer air into the cloud resulted in low ozone throughout the tower and the anvil. Stratospheric air with high-ozone mixing ratios was brought into the upper portion of the anvil. The model did not show any significant transport of boundary layer gases into the stratosphere.

Regional

Regional estimates of deep convective transport have been made through use of a traveling one-dimensional model, regional transport models driven by parameterized convective mass fluxes from mesoscale meteorological models, and a statistical-dynamical approach. Chatfield and Delany (1990) simulated convective transport for a hypothetical case over South America during the biomass burning season using a traveling one-dimensional model containing cloud-scale vertical transport and chemistry. They showed that the "mix-then-cook" scenario of rapid vertical transport of ozone precursors in deep convection allowed a more persistent increase in the

tropospheric ozone column over a wide region than did the "cook-then-mix" scenario of transport in a fair weather boundary layer for several days prior to venting.

Pickering et al. (1992c) used a combination of deep convective cloud cover statistics from the International Satellite Cloud Climatology Project (ISCCP) and convective transport statistics from GCE model simulations of prototype storms to estimate that between 10 and 40% of CO from biomass burning in the Brazilian state of Rondonia is vented from the boundary layer by deep convection. The statistical-dynamical approach was also used by Thompson et al. (1994) to estimate the convective transport component of the boundary layer CO budget for the central United States for the month of June (Fig. 9). Deep convective venting of the boundary layer dominated other components of the CO budget during early summer, providing a net (upward minus downward) flux of 18.1×10^8 kg CO/month to the free troposphere. In this respect the central United States acts as a "chimney" for the country.

Regional chemical transport models (CTMs) have been used for applications such as simulations of photochemical ozone production, acid deposition, and fine particulate matter. Walcek et al. (1990) included a parameterization of cloud-scale

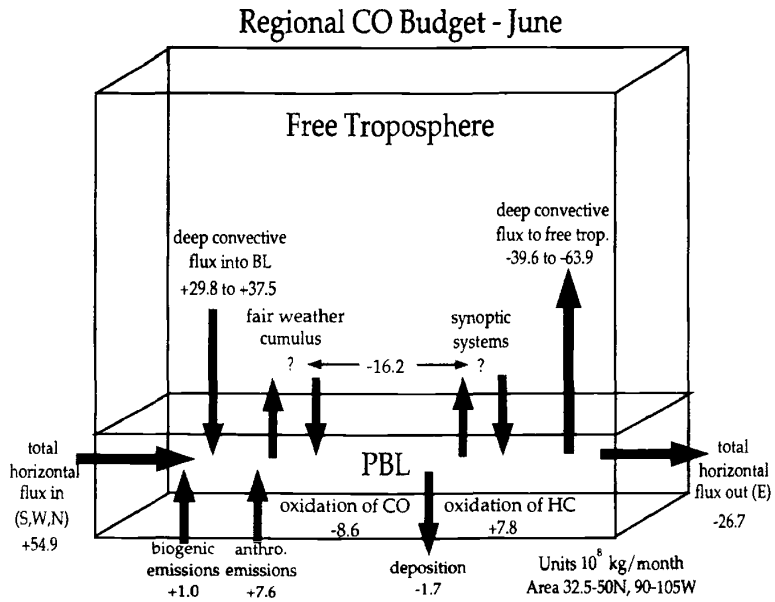


Figure 9 Regional boundary layer CO budget for the central United States (32.5°N to 50°N; 90°N to 105°W). Note magnitudes of upward and downward deep convective transport components. Question marks signify that relative amounts of CO flux due to shallow convection and synoptic-scale systems are unknown. From Thompson et al. (1994).

aqueous chemistry, scavenging, and vertical mixing in the chemistry model of Chang et al. (1987). The vertical distribution of cloud microphysical properties and the amount of subcloud-layer air lifted to each cloud layer are determined using a simple entrainment hypothesis (Walcek and Taylor, 1986). Vertically integrated O_3 formation rates over the northeast United States were enhanced by $\sim 50\%$ when the in-cloud vertical motions were included in the model.

Wang et al. (1996) simulated the September 26–27, 1992, TRACE–A mesoscale convective systems (MCS) and the June 10–11, 1985, PRE-STORM squall line with the NCAR/Penn State Mesoscale Model (MM5; Grell et al., 1994; Dudhia, 1993). Convection is parameterized as a subgrid-scale process in MM5; two convective parameterizations were tested in the Wang et al. (1996) work. These were the Grell (1993) and Kain and Fritsch (1993) schemes. Mass fluxes and detrainment profiles from these schemes were used along with the three-dimensional wind fields in CO tracer transport calculations for the two convective events. The time-evolving tracer fields in the upper troposphere are different in the tropical MCS and the midlatitude squall line. The nearly stationary tropical system produced regions of large upper tropospheric CO that moved very little in the horizontal by the end of the 24-h simulation, whereas enhanced upper tropospheric CO propagates with the relatively fast moving midlatitude squall line. Using a grid size of 25 to 30 km, the parameterized subgrid vertical transport represented 48% (Grell, 1993) and 41% (Kain and Fritsch, 1993) of the total upward transport in the tropical case and 64% (Kain–Fritsch) in the midlatitude case. Pickering et al. (1996) demonstrated that the MM5 convective transport (Fig. 10) reproduced the observed factor of three enhancement of upper tropospheric CO and that over several days downwind transport the enhanced upper tropospheric O_3 precursor mixing ratios allowed O_3 production to proceed at a rate ~ 4 times faster than would have occurred in undisturbed air. The U.S. Environmental Protection Agency (EPA) has developed a Community Multiscale Air Quality (CMAQ) modeling system that uses MM5 with the Kain–Fritsch convective scheme as the dynamical driver (Ching et al., 1998).

Global

Convective transport in global chemistry and transport models is treated as a subgrid-scale process that is parameterized typically using cloud mass flux information from a general circulation model (GCM) or global data assimilation system. Jacob and Prather (1990) simulated the distribution of radon-222 over North America using a three-dimensional CTM driven with meteorological fields from the NASA Goddard Institute for Space Studies (GISS) GCM II (Hansen et al., 1983), having a horizontal resolution of $4^\circ \times 5^\circ$ and 9 layers in the vertical. Simulation of convective transport in the CTM follows the scheme used in the GCM to transport momentum, sensible heat, and moisture. The model gave a reasonable simulation of radon-222 observations over the United States, but with some significant discrepancies that were traced to problems in the GCM meteorology. Improved simulations of transport have been obtained using a newer convective parameterization of Del Genio and Yao (1988).

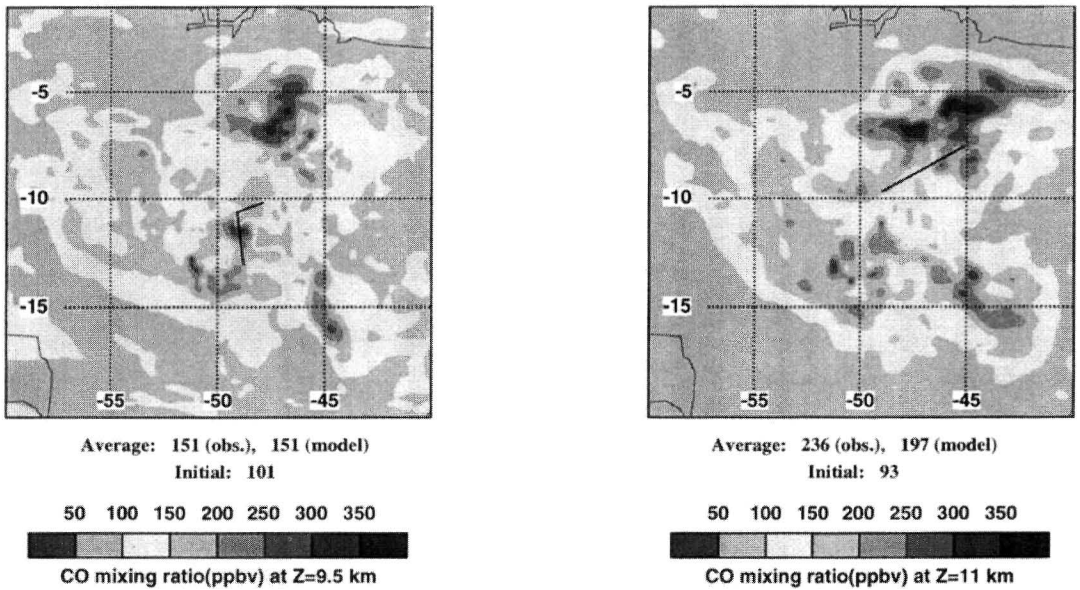


Figure 10 MM5 simulation result for CO tracer following TRACE-A mesoscale convective events. Shown are CO mixing ratios at 1200 UT September 27, 1992, at altitudes 9.5 and 11 km. Region shown is fine-grid (30-km resolution) domain of MM5 simulation. Includes grid-scale and subgrid transport. From Pickering et al. (1996).

While GCMs can provide data only for a "typical" year, data assimilation systems can provide "real" day-by-day meteorological conditions, such that CTM output can be compared directly with observations of trace gases. The NASA Goddard Earth Observing System Data Assimilation System (GEOS-1 DAS; Schubert et al., 1993) provides archived global data sets for the period 1980–1995, at $2° \times 2.5°$ resolution with 20 layers in the vertical. Convection is parameterized with the relaxed Arakawa–Schubert scheme (Moorthi and Suarez, 1992). Pickering et al. (1995) showed that the cloud mass fluxes from GEOS-1 DAS are reasonable for the June 10–11, 1985, PRE-STORM squall line based on comparisons with the GCE model (cloud-resolving model) simulations of the same storm (Fig. 11). In addition, the GEOS-1 DAS cloud mass fluxes compared favorably with the regional estimates of convective transport for the central United States presented by Thompson et al. (1994). Allen et al. (1996a,b) have used the GEOS-1 DAS data to drive global CTM calculations for radon-222 and for CO. However, Allen et al. (1997) have shown that the GEOS-1 DAS overestimates the amount and frequency of convection in the tropics and underestimates the convective activity over midlatitude marine storm tracks.

Mahowald et al. (1995) investigated the behavior of seven different cumulus parameterization schemes in deriving convective transport from meteorological analysis data sets that did not routinely archive cloud mass fluxes. The derived convective transport was used in a column model and showed that the resulting vertical profile of trace gases was highly sensitive to the parameterization used.

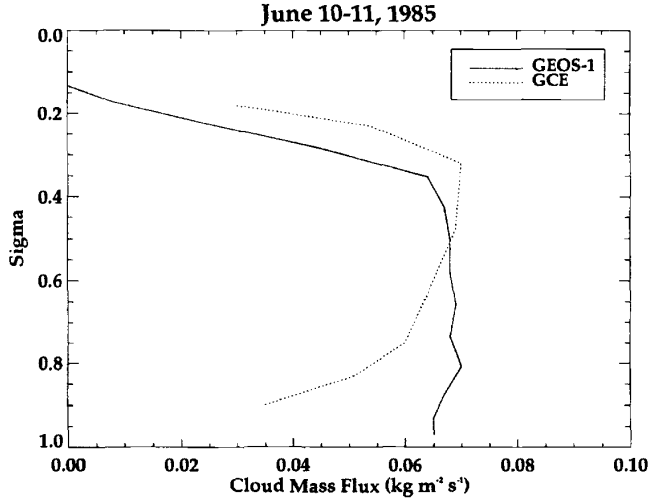


Figure 11 Profiles of cloud mass flux for June 10–11, 1985, PRE-STORM squall line computed by GEOS-1 DAS and by the GCE model. From Pickering et al. (1995).

Rasch et al. (1997) have described use of the output from the NCAR (National Center for Atmospheric Research) Community Climate Model (CCM3) in a chemical transport model. This CTM uses results from the CCM3 convective parameterizations [(Zhang and McFarlane (1995) penetrative convection parameterization and the Hack (1994) scheme for shallow convection)].

4 SUMMARY

Observations and model simulations over the last 15 years have greatly clarified the role of convection in transporting trace constituents in the atmosphere. It is now well established that some nonprecipitating cumulus clouds aid in venting the boundary layer. However, methods to determine the fraction of such clouds that actively transport trace gases to the free troposphere for a region on a given day still require further work. It is also well established that deep convection can transport large quantities of boundary layer gases to the middle and upper troposphere where they have a much longer chemical lifetime and can be transported large distances from their source region. Ozone production in the free troposphere can be enhanced by a factor of 4 or more as a result of deep convection. Downdrafts in convective storms can transport cleaner air from the midtroposphere down to the boundary layer. In remote regions low values of ozone and NO_x can be transported to the upper troposphere, decreasing ozone and ozone production at these altitudes in such regions. In addition, convection induces downward transport of larger O_3 mixing ratios into the remote boundary layer where photochemistry and surface deposition destroy O_3. Storms that reach near or above the preconvective tropopause can induce exchange of trace constituents between stratosphere and troposphere.

Cloud-resolving models are the best tool for detailed studies of convective transport by individual storm systems. Air parcel trajectories and tracer transport calculations using the wind fields from such models are useful for understanding the flow patterns involved in the convective transport process. A cloud model is also useful in evaluating parameterized convective transport in regional or global models. Considerable uncertainty still exists in the output of convective parameterizations concerning the frequency, location, and magnitude of vertical transport, making convective transport one of the largest sources of uncertainty in regional and global CTMs.

REFERENCES

Allen, D. J., P. Kasibhatla, A. M. Thompson, R. B. Rood, B. G. Doddridge, K. E. Pickering, R. D. Hudson, and S.-J. Lin, Transport-induced interannual variability of carbon monoxide determined using a chemistry and transport model, *J. Geophys. Res.*, *101*, 28655–28669, 1996a.

Allen, D. J., R. B. Rood, A. M. Thompson, and R. Hudson, Three dimensional radon 222 calculations using assimilated meteorological data and a convective mixing algorithm, *J. Geophys. Res.*, *101*, 6871–6881, 1996b.

Allen, D. J., K. E. Pickering, and A. Molod, An evaluation of deep convective mixing in the Goddard chemical transport model using ISCCP cloud parameters, *J. Geophys. Res.*, *102*, 25467–25476, 1997.

Chang, J. S., R. A. Brost, I. S. I. Isaksen, S. Madronick, P. Middleton, W. R. Stockwell, C. J. Walcek, A three-dimansional Eulerian acid deposition model: Physical concepts and formulation, *J. Geophys. Res.*, *92*, 14, 681–14, 700, 1987.

Chatfield, R. B., and P. J. Crutzen, Sulfur dioxide in remote oceanic air: Cloud transport of reactive precursors, *J. Geophys. Res.*, *89*, 7111–7132, 1984.

Chatfield R. B., and A. C. Delany, Convection links biomass burning to increased tropical ozone: However, models will tend to overpredict O_3, *J. Geophys. Res.*, *95*, 18473–18488, 1990.

Ching, J. K. S., and A. J. Alkezweeny, Tracer study of vertical exchange by cumulus clouds, *J. Clim. Appl. Meteorol.*, *25*, 1702–1711, 1986.

Ching, J. K. S., S. T. Shipley, and E. V. Browell, Evidence for cloud venting of mixed layer ozone and aerosols, *Atmos. Environ.*, *22*, 225–242, 1988.

Ching, J. K. S., D. W. Byun, J. Young, F. Binkowski, J. Pleim, S. Roselle, J. Godowitch, W. Benjey, and G. Gipson, Science features in Models-3 Community Multiscale Air Quality System, in *Preprints of the Tenth Joint AMS/AWMA Conference on Applications of Air Pollution Meteorology*, Phoenix, AZ, 1998.

Crutzen, P. J., and L. T. Gidel, A two-dimensional photochemical model of the atmosphere, 2. The tropospheric budgets of the anthropogenic chlorocarbons, CO, CH_4, CH_3Cl and the effect of various NO_x sources on tropospheric ozone, *J. Geophys. Res.*, *88*, 6641–6661, 1983.

Danielsen, E. F., In-situ evidence of rapid, vertical, irreversible transport of lower tropospheric air into the lower tropical stratosphere by convective cloud turrets and by large-scale upwelling in tropical cyclones, *J. Geophys. Res.*, *98*, 8665–8681, 1993.

Del Genio, A. D., and M. S. Yao, Sensitivity of a global climate model to the specification of convective updraft and downdraft mass fluxes, *J. Atmos. Sci.*, *45*, 2641–2668, 1988.

Dickerson, R. R., G. J. Huffman, W. T. Luke, L. J. Nunnermacker, K. E. Pickering, A. C. D. Leslie, C. G. Lindsey, W. G. N. Slinn, T. J. Kelly, P. H. Daum, A. C. Delany, J. P. Greenberg, P. R. Zimmerman, J. F. Boatman, J. D. Ray, and D. H. Stedman, Thunderstorms: An important mechanism in the transport of pollutants, *Science*, *235*, 460–465, 1987.

Dudhia, J., A nonhydrostatic version of the Penn State/NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front, *Monthly Weather Rev.*, *121*, 1493–1513, 1993.

Garstang, M., J. Scala, S. Greco, R. Harriss, S. Beck, E. Browell, G. Sachse, G. Gregory, G. Hill, J. Simpson, W.-K. Tao, and A. Torres, Trace gas exchanges and convective transports over the Amazonian rain forest, *J. Geophys. Res.*, *93*, 1528–1550, 1988.

Gidel, L. T., Cumulus cloud transport of transient tracers, *J. Geophys. Res.*, *88*, 6587–6599, 1983.

Greenhut, G. K., Transport of ozone between boundary layer and cloud layer by cumulus clouds, *J. Geophys. Res.*, *91*, 8613–8622, 1986.

Greenhut, G. K., J. K. S. Ching, R. Pearson, Jr., and T. P. Repoff, Transport of ozone by turbulence and clouds in an urban boundary layer, *J. Geophys. Res.*, *89*, 4757–4766, 1984.

Grell, G. A., Prognostic evaluation of assumptions used by cumulus parameterizations, *Monthly Weather Rev.*, *121*, 764–767, 1993.

Grell, G. A., J. Dudhia, and D. Stauffer, *A Description of the Fifth Generation Penn State/NCAR Mesoscale Model (MM5)*, NCAR/TN-389 + STR, National Center for Atmospheric Research, Boulder, CO, 1994.

Hack, J. J., Parameterization of moist convection in the NCAR Community Climate Model, CCM2, *J. Geophys. Res.*, *99*, 5551–5568, 1994.

Hansen, J., G. Russell, D. Rind, P. Stone, A. Lacis, S. Lebendeff, R. Ruedy, and L. Travis, Efficient three-dimensional global models for climate studies: Models I and II, *Monthly Weather Rev.*, *111*, 609–662, 1983.

Jacob, D. J., and M. J. Prather, Radon-222 as a test of convective transport in a general circulation model, *Tellus*, *42B*, 118–134, 1990.

Kain, J. S., and J. M. Fritsch, Convective parameterization in mesoscale models: The Kain-Fritsch scheme, in K. A. Emanuel and D. J. Raymond (Eds.), *The Representation of Cumulus Convection in Numerical Models*, American Meteorological Society, Boston, MA, 1993.

Kawakami, S., Y. Kondo, M. Koike, H. Nakajima, G. L. Gregory, G. W. Sachse, R. E. Newell, E. V. Browell, D. R. Blake, J. M. Rodriguez, and J. T. Merrill, Impact of lightning and convection on reactive nitrogen in the tropical free troposphere, *J. Geophys. Res.*, *102*, 28367–28384, 1997.

Kleinman, L. I., and P. H. Daum, Vertical distribution of aerosol particles, water vapor, and insoluble trace gases in convectively mixed air, *J. Geophys. Res.*, *96*, 991–1005, 1991.

Kley, D., P. J. Crutzen, H. G. J. Smit, H. Vomel, S. J. Oltmans, H. Grassl, and V. Ramanathan, Observations of near-zero ozone concentrations over the convective Pacific: Effects on air chemistry, *Science*, *274*, 230–233, 1996.

Lelieveld, J., and P. J. Crutzen, Role of deep cloud convection in the ozone budget of the troposphere, *Science*, *264*, 1759–1761, 1994.

Luke, W. T., R. R. Dickerson, W. F. Ryan, K. E. Pickering, and L. J. Nunnermacker, Tropospheric chemistry over the lower Great Plains of the United States 2. Trace gas profiles and distributions, *J. Geophys. Res.*, *97*, 20647–20670, 1992.

Mahowald, N. M., P. J. Rasch, and R. G. Prinn, Cumulus parameterizations in chemical transport models, *J. Geophys. Res.*, *100*, 26173–26190, 1995.

Moorthi, S., and M. J. Suarez, Relaxed Arakawa-Schubert: A parameterization of moist convection for general circulation models, *Monthly Weather Rev.*, *120*, 978–1002, 1992.

Newell, R. E., et al., Atmospheric sampling of Supertyphoon Mireille with NASA DC-8 aircraft on September 27, 1991, during PEM-West A, *J. Geophys. Res.*, *101*, 1853–1871, 1996.

Pickering, K. E., R. R. Dickerson, G. J. Huffman, J. F. Boatman, and A. Schanot, Trace gas transport in the vicinity of frontal convective clouds, *J. Geophys. Res.*, *93*, 759–773, 1988.

Pickering, K. E., R. R. Dickerson, W. T. Luke, and L. J. Nunnermacker, Clear-sky vertical profiles of trace gases as influenced by upstream convective activity, *J. Geophys. Res.*, *94*, 14879–14892, 1989.

Pickering, K. E., A. M. Thompson, R. R. Dickerson, W. T. Luke, and D. P. McNamara, Model calculations of tropospheric ozone production potential following observed convective events, *J. Geophys. Res.*, *95*, 14049–14062, 1990.

Pickering, K. E., A. M. Thompson, J. R. Scala, W.-K. Tao, J. Simpson, and M. Garstang, Photochemical ozone production in tropical squall line convection during NASA Global

Tropospheric Experiment/Amazon Boundary Layer Experiment 2A, *J. Geophys. Res.*, *96*, 3099–3114, 1991.

Pickering, K. E., A. M. Thompson, J. Scala, W.-K. Tao, R. R. Dickerson, and J. Simpson, Free tropospheric ozone production following entrainment of urban plumes into deep convection, *J. Geophys Res.*, *97*, 17985–18000, 1992a.

Pickering, K. E., A. M. Thompson, J. R. Scala, W.-K. Tao, and J. Simpson, Ozone production potential following convective redistribution of biomass emissions, *J. Atmos. Chem.*, *14*, 297–313, 1992b.

Pickering, K. E., A. M. Thompson, W.-K. Tao, and T. L. Kucsera, Upper tropospheric ozone production following mesoscale convection during STEP/EMEX, *J. Geophys. Res.*, *98*, 8737–8749, 1993.

Pickering, K. E., A. M. Thompson, W.-K. Tao, R. B. Rood, D. P. McNamara, and A. M. Molod, Vertical transport by convective clouds: Comparisons of three modeling approaches, *Geophys. Res. Lett.*, *22*, 1089–1092, 1995.

Pickering, K. E., J. R. Scala, A. M. Thompson, W.-K. Tao, and J. Simpson, A regional estimate of the convective transport of CO from biomass burning, *Geophys. Res. Lett.*, *19*, 289–292, 1992c.

Pickering, K. E., A. M. Thompson, Y. Wang, W.-K. Tao, D. P. McNamara, V. W. J. H. Kirchhoff, B. G. Heikes, G. W. Sachse, J. D. Bradshaw, G. L. Gregory, and D. R. Blake, Convective transport of biomass burning emissions over Brazil during TRACE-A, *J. Geophys. Res.*, *101*, 23993–24012, 1996.

Piotrowicz, S. R., H. F. Bezdek, G. R. Harvey, M. Springer-Young, and K. J. Hanson, On the ozone minimum over the equatorial Pacific Ocean, *J. Geophys. Res.*, *96*, 18679–18687, 1991.

Poulida, O., R. R. Dickerson, and A. Heymsfield, Troposphere-stratosphere exchange in a midlatitude mesoscale convective complex: 1. Observations, *J. Geophys. Res.*, *101*, 6823–6836, 1996.

Rasch, P. J., N. M. Mahowald, and B. E. Eaton, Representations of transport, convection, and the hydrologic cycle in chemical transport models: Implications for the modeling of short-lived and soluble species, *J. Geophys. Res.*, *102*, 28127–28152, 1997.

Riehl, H., and J. Simpson, The heat balance of the equatorial zone, revisited, *Beitr. Phys. Atmos.*, *52*, 287–305, 1979.

Scala, J., M. Garstang, W.-K. Tao, K. Pickering, A. Thompson, J. Simpson, V. Kirchhoff, E. Browell, G. Sachse, A. Torres, G. Gregory, R. Rasmussen, and M. Khalil, Cloud draft structure and trace gas transport, *J. Geophys. Res.*, *95*, 17017–17030, 1990.

Schubert, S. D., R. B. Rood, and J. Pfaendtner, An assimilated data set for earth science applications, *Bull. Am. Meteoral. Soc.*, *74*, 2331–2342, 1993.

Stenchikov, G., R. Dickerson, K. Pickering, W. Ellis, B. Doddridge, S. Kondragunta, and O. Poulida, Stratosphere-troposphere exchange in a mid-latitude mesoscale convective complex: Part 2, Numerical simulations, *J. Geophys. Res.*, *101*, 6837–6851, 1996.

Stull, R. B., A fair-weather cumulus cloud classification scheme for mixed layer studies, *J. Clim. Appl. Meteorol.*, *24*, 49–56, 1985.

Suhre, K., J.-P. Cammas, P. Nedelec, R. Rosset, A. Marenco, and H. G. J. Smit, Ozone-rich transients in the upper equatorial Atlantic troposphere, *Nature*, *388*, 661–663, 1997.

Tao, W.-K., and J. Simpson, The Goddard cumulus ensemble model. Part I: Model description, *Terrestr. Atmos. Oceanic Sci.*, *4*, 35–72, 1993.

Thompson, A. M., K. E. Pickering, R. R. Dickerson, W. G. Ellis, Jr., D. J. Jacob, J. R. Scala, W.-K. Tao, D. P. McNamara, and J. Simpson, Convective transport over the Central United States and its role in the regional CO and O_3 budgets, *J. Geophys. Res.*, *99*, 18, 703–18, 711, 1994.

Thompson, A. M., W.-K. Tao, K. E. Pickering, J. R. Scala, and J. Simpson, Tropical deep convection and ozone formation, *Bull. Am. Meteorol. Soc.*, *78*, 1043–1054, 1997.

Vukovich, F. M., and J. K. S. Ching, A semi-impirical approach to estimate vertical transport by nonprecipitating convective clouds on a regional scale, *Atmos. Environ.*, *24A*, 2153–2168, 1990.

Walcek, C. J., W. R. Stockwell, and J. S. Chang, Theoretical estimates of the dynamic, radiative, and chemical effects of clouds on tropospheric trace gases, *Atmos. Res.*, *25*, 53–69, 1990.

Walcek, C. J., and G. R. Taylor, A theoretical method for computing vertical distribution of acidity and sulfate production within cumulus clouds, *J. Atmos. Sci.*, *43*, 339–355, 1986.

Wang, C., and J. S. Chang, A three-dimensional numerical model of cloud dynamics, microphysics, and chemistry 1, Concepts and formulation, *J. Geophys. Res.*, *98*, 14827–14844, 1993.

Wang, C., P. J. Crutzen, V. Ramanathan, and S. F. Williams, The role of a deep convective storm over the tropical Pacific Ocean in the redistribution of atmospheric chemical species, *J. Geophys. Res.*, *100*, 11509–11516, 1995.

Wang, Y., W.-K. Tao, K. E. Pickering, A. M. Thompson, J. S. Kain, R. F. Adler, J. Simpson, P. R. Keehn, and G. S. Lai, Mesoscale model simulations of TRACE-A and PRE-STORM convective systems and associated tracer transport, *J. Geophys. Res.*, *101*, 24013–24027, 1996.

Zhang, G. J., and N. A. McFarlane, Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model, *Atmos. Ocean*, *33*, 407–446, 1995.

CHAPTER 10

BOUNDARY LAYER PROCESSES AND FLUX MEASUREMENTS

DONALD H. LENSCHOW

1 INTRODUCTION

The planetary boundary layer (PBL) is that part of the atmosphere that interacts with Earth's surface on a time scale of about an hour. This rapid interaction is a direct result of turbulence, which is an essential feature of the PBL. The sources of turbulence are wind shear and convection. Defining characteristics of turbulence are its chaotic fluctuations and diffusiveness. That is, trace constituents released into a turbulent fluid are rapidly diffused, and the small-scale patterns of this diffusion cannot be predicted. Because of the randomness and the large range of scales of PBL turbulence, processes in the PBL are often described in terms of statistical averages of fluctuations. This means that most measurements of PBL structure need to be spatially or temporally averaged before they can be quantitatively interpreted.

If generation of turbulence by convection is occurring in the PBL, it is known as an *unstable* or *convective* boundary layer (CBL); if the hydrodynamic stratification of the PBL acts to suppress or dissipate turbulence, it is known as a *stable* boundary layer (SBL). When relative humidity reaches 100% within the PBL, clouds form that can have a dramatic effect on its subsequent evolution.

2 BOUNDARY LAYER EVOLUTION

Over land, the daily solar cycle determines the PBL evolution. In the morning, the sun starts to warm the ground, which has been cooling through the night by infrared radiation. Clear air is nearly transparent to the sun's short-wave (visible) radiation

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

and thus is warmed only slightly by direct solar radiation. Instead, the ground absorbs most of the solar radiation and then warms the air above it mostly through convection, which is the upward movement of buoyant parcels of air warmed by contact with the surface, in combination with compensating downward transfer of cooler, more dense air from above. This process generates turbulence that, in turn, increases the efficiency of transport of atmospheric constituents in the CBL. Efficient mixing means that the *lapse rate*, which is the rate of change of temperature with height, is nearly adiabatic throughout much of the CBL. This means that vertical displacements of an air parcel do not change the buoyancy of the parcel relative to its environment. Figure 1 shows the structure of the CBL using *virtual potential temperature*, which is constant with height in an adiabatic layer, and a scalar variable with a surface source and negligible concentration above the CBL.

The CBL continues to deepen typically at least until early afternoon to perhaps 1 to 3 km. Generally, the *relative humidity* in the upper part of the CBL tends to

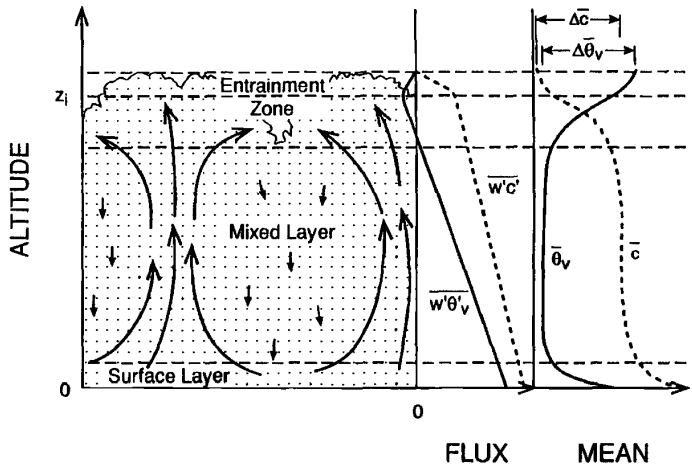


Figure 1 Convective boundary layer. On the left, the sublayers that make up the boundary layer are shown, along with a schematic of flow patterns—upward-moving buoyant thermals forming near the top of the surface layer, extending through the mixed layer, and dissipating in the upper part of the boundary layer. Part of their kinetic energy is dissipated in the entrainment zone by entraining warmer, more buoyant air from above into the boundary layer. The thermals have a smaller total area, and consequently a larger velocity magnitude than the compensating downward-moving air in between the thermals. The flux profiles in the middle panel show the virtual potential temperature (θ_v) flux (which has a universal shape), and the flux of a scalar $\mathscr{C}$ which (in this case) has a source at the surface and whose mean concentration decreases with height throughout and immediately above the boundary layer. The virtual potential temperature can normally be assumed to be a conserved variable in a well-mixed clear boundary layer. The mean virtual potential temperature and scalar concentration profiles, including their jumps across the top of the boundary layer, are shown on the right.

increase through the day from moistening due to surface *evapotranspiration* and turbulent mixing. If the humidity reaches saturation, clouds develop at the top of the CBL.

As the solar heating decreases late in the day, convection disappears and radiative cooling at the surface again dominates over solar warming. Eventually, the ground becomes cooler than the overlying air, which radiatively cools much more slowly than the ground. At this point, the SBL develops, which is considerably shallower (ranging from a few tens to a few hundred meters deep) and less turbulent because buoyancy is now suppressing the turbulence generated by shear. As a result, turbulent transport is less efficient. Above the surface layer, turbulence becomes intermittent and, because of the stable stratification, gravity waves may become important. The top of the SBL is not as well defined as the daytime CBL since the turbulence decreases much more slowly with height. Because of the low turbulence, discrete layers with their own characteristic properties may form and advect at different speeds. After sunrise, the ground begins to warm, and the cycle repeats.

Over the ocean, the large heat capacity and effective heat conductivity keep the ocean temperature nearly constant over the daily cycle; thus the daily cycle is often insignificant in the marine boundary layer (MBL). This means that the MBL typically has much less production of turbulence energy by buoyancy and is usually shallower (perhaps 0.5 to 1.5 km) than the daytime CBL over land. On average, there is less cloudiness at the top of the MBL during the day than at night because absorption of solar short-wave radiation warms the MBL directly and thus reduces the production of turbulence by buoyancy and the degree of mixing.

3 STRUCTURE OF THE BOUNDARY LAYER

Boundary layer processes are dominated by the diffusive character of turbulence. This diffusion can be described as a flux of a constituent, which is the rate of transport of the constituent across a surface per unit time and per unit area. Alternatively, it can be expressed as a constituent density times a velocity. Both velocity components and scalars are defined as sums of a mean and a fluctuation, $\mathscr{U}_i = U_i + u_i$ and $\mathscr{S} = S + s$. The horizontal wind components are commonly defined as $\mathscr{U}_1 \equiv \mathscr{U}$ and $\mathscr{U}_2 \equiv \mathscr{V}$, while the vertical component is $\mathscr{U}_3 \equiv \mathscr{E}$). The average of the fluctuations, $\bar{u}_i = \bar{s} = 0$, where the overbar is the common convention to denote an average of a turbulence variable over a time period or a length long enough to give a stable estimate of its mean.

In a turbulent fluid, the scalar flux is defined as the sum or integral of fluctuations in the velocity normal to a surface times the concurrent fluctuations in constituent density divided by the time period or length over which the sum or integral is calculated. Normally in the PBL, only the vertical component of the flux is of interest. Thus,

$$F_s = \frac{1}{T}\int_0^T ws\, dt = \overline{ws} \tag{1}$$

where w and s are fluctuations in the vertical wind component and a scalar density, respectively, and T is the averaging time period. Equation (1) defines the *eddy correlation* technique for measuring flux.

The PBL can be further broken down into sublayers: The lowest few tens of meters is called the *surface layer*. In this region, which is less than 10% of the depth of the PBL, the fluxes can be considered constant with height, but for variables with large surface fluxes, the mean vertical gradients of these variables are large relative to the rest of the PBL. It is common to relate the flux in the surface layer to a gradient by a diffusivity,

$$F_s = -K_s \frac{\partial S}{\partial z} \tag{2}$$

Since transport by turbulent eddies in the surface layer is roughly about 10^5 times more efficient than transport by molecular diffusion, for scalars we call K_s the eddy diffusivity and for momentum the eddy viscosity. Near the surface, the typical maximum horizontal dimension of the eddies making important contributions to the flux is roughly about 200 times the height above the surface. Since the efficiency of turbulent transport scales with the size of the eddies, the eddy diffusivity increases approximately linearly with height. Equivalently, since the flux is approximately constant in the surface layer, the gradient decreases approximately inversely with height near the surface.

In the CBL, the layer above the surface layer and below the region near the PBL top is called the *mixed layer*. Since this encompasses the bulk of the PBL, the fluxes show considerable variability. Here the gradients are small because the mixing process is efficient. The individual turbulent eddies extend throughout the mixed layer and are called *thermals* (or *plumes*). The typical maximum size of eddies making important contributions to the flux changes more slowly with height than in the surface layer, and is roughly about 40 times the depth of the CBL. The turbulence energy (the sum of the three component velocity variances) reaches a maximum at about one third the height of the mixed layer.

Above the mixed layer is the *entrainment layer*. In this region, a sharp interface typically occurs between the CBL and the overlying nonturbulent *free atmosphere*. This interface is generated by turbulent eddies that protrude into the nonturbulent layer, engulf or capture volumes of nonturbulent air, and fall back into the mixed layer. Smaller-scale turbulence within these larger eddies then commingles this entrained air with CBL air so that it becomes part of the mixed layer. Details of the clear (cloud-free) CBL structure are shown in Figure 1.

If clouds form at the top of the CBL, one of two scenarios occurs: If the density decrease across the top of the CBL is small enough that condensation and mixing processes decrease the density to a value below the overlying air, the cloud can penetrate through the top of the CBL and form a *cumulus*. In that case, the CBL top is approximately at cloud base, and the clouds penetrate into the overlying air until mixing with their environment limits their growth and they lose their buoyancy. This *venting* process injects CBL air into the overlying atmosphere, which is a way to

increase humidity above the CBL and introduce trace constituents originating at the surface or within the CBL into the free atmosphere. Compensating downward motion can dilute the concentration of pollutants in the CBL. If the decrease in density across the top of the CBL is large enough that phase changes and mixing of overlying air with CBL air do not decrease the density sufficiently to lower the density below the overlying air, the cloud layer is contained within the CBL and a *stratus* or *stratocumulus* cloud layer may exist. In this case, the CBL maintains a sharp interface. Radiative cooling at cloud top can also generate CBL turbulence and contribute to the efficiency of mixing in the cloud-capped CBL.

In the SBL, the layer above the surface layer is a region of decreasing intensity and increasing intermittency of turbulence. Here the individual turbulent eddies may not extend throughout the SBL, but occur in sublayers that develop locally enhanced shear that may intermittently break down into turbulence. The turbulence is then dissipated with the net result that the gradients in the sublayer are reduced by the transient turbulence event. The process may very well be repeated over time. Since the turbulence is more local in nature, the scales and structure are less well defined than in the CBL, and velocity and scalar variances decrease with height throughout the SBL. Relatively large gradients of both wind and scalars, and multiple sublayers that are only intermittently coupled may exist. The top of the SBL generally does not have a well-defined lid. Because of the intermittency and smaller length scales of the mixing process, the shallower and less well defined structure of the SBL, and the presence of gravity waves that produce velocity fluctuations but no flux, trace constituent fluxes are much more problematic and much less frequently measured in the SBL than the CBL. For these reasons, most of the subsequent discussion deals solely with the CBL.

4 SCALES AND PROCESSES

Wind shear is the rate of change of wind with height,

$$\frac{\partial U}{\partial z} + \frac{\partial V}{\partial z} \tag{3}$$

where U and V are the averaged horizontal wind components. Because of drag induced by Earth's surface, the horizontal wind approaches zero at the surface. Since the eddy viscosity increases approximately linearly with height and the kinematic momentum flux ($\overline{uw}$) is approximately constant [and equal to $(\overline{uw})_0$] in the surface layer, the wind shear decreases roughly inversely with height very near the surface. Further above the surface, the wind shear, as well as scalar gradients in the surface layer depend also on the stability—that is, if the surface buoyancy flux is positive, the magnitudes of the wind shear and scalar gradients decrease with height less rapidly than the inverse of height, and if the buoyancy flux is negative, they decrease more rapidly than the inverse of height. Furthermore, the direction of the mean wind is assumed to be constant with height in the surface layer, so the

coordinate system can be defined such that $V = 0$. Thus near the surface, or in a surface layer with zero surface buoyancy flux (a neutrally stratified PBL),

$$\frac{\partial U}{\partial z} = \frac{u_*}{kz} \tag{4}$$

where $u_*^2 = -(\overline{uw})_0$ is the friction velocity and k is the von Kármán constant ($\simeq 0.4$). A typical range of values for u_* over a treeless vegetated surface in moderate winds would be 0.2 to 0.8 m/s. A similar relation holds for scalar quantities in the surface layer,

$$\frac{\partial S}{\partial z} = \frac{S_*}{kz} \tag{5}$$

where $S_* = -F_{s0}/u_*$ and F_{s0} is the surface-layer flux of $\mathscr{S}$.

These equations can be integrated to obtain vertical profiles,

$$U(z_2) - U(z_1) = \frac{u_*}{k} \ln \frac{z_2}{z_1} \tag{6}$$

and

$$S(z_2) - S(z_1) = \frac{S_*}{k} \ln \frac{z_2}{z_1} \tag{7}$$

Since $U \to 0$ at the surface, we define the *roughness length* z_0 as the height at which the extrapolated wind profile goes to zero so that

$$U(z) = \frac{u_*}{k} \ln \frac{z}{z_0} \tag{8}$$

The roughness length is approximately $\frac{1}{30}$ the height of individual surface roughness elements. It ranges from about 10^{-4} m over calm water to about 0.5 m over a forest.

The production of turbulence energy by wind shear is given by the product of mean wind shear and (kinematic) momentum flux,

$$E_u = -\overline{uw}\left(\frac{\partial U}{\partial z} + \frac{\partial V}{\partial z}\right) \tag{9}$$

Near the surface, or in a neutral PBL, inserting (4) into (9) reduces (9) to

$$E_u = \frac{u_*^3}{kz} \tag{10}$$

Production (dissipation) of turbulence by convection can be expressed in terms of a buoyancy flux,

$$F_b = \frac{g}{T}\overline{wT_v} \tag{11}$$

where g is gravity, T is temperature, and T_v is virtual temperature, which includes the density effects of water vapor on the temperature. The parameter g/T is the buoyancy parameter, which is the thermal expansion coefficient of air times the acceleration of gravity. Since the buoyancy flux can also be negative, this term may also act to dissipate turbulence. The total turbulence energy production is the sum of (11) and (9). Averaged over the entire PBL, the energy production must be equal to the turbulence dissipation, which is the loss of turbulence energy due to the viscous forces that occur predominantly at very small scales (i.e., less than a few centimeters). In effect, the viscous forces convert the kinetic energy of turbulence into thermal energy, and thus heat the air (although the temperature increase is insignificant).

Near the surface, the negative ratio of energy production by buoyancy to production by shear is given by

$$-\frac{F_{b0}}{u_*^3/kz} \tag{12}$$

The length at which this ratio is unity, called the *Obukhov length*, is given by

$$L = -\frac{u_*^3}{kF_{b0}} \tag{13}$$

This is a measure of the stability of the surface layer and is used as a scaling height to normalize the observation height. Similarly, velocity, temperature, and scalar variables in the surface layer can be normalized by u_*, $T_* = -(\overline{wT})_0/u_*$, and S_*. In this way, normalized surface layer variables as functions of height can be expressed as universal functions in both the unstably and stably stratified surface layer. This is a powerful technique for relating surface layer measurements to a universal surface layer structure in diabatic (nonzero surface buoyancy flux) PBLs with enough mean wind to generate a well-defined u_*. For example, (4) and (5) can be extended to the diabatic surface layer by including stability functions in the formulations,

$$\frac{\partial U}{\partial z} = \frac{u_*}{kz}\phi_m\left(\frac{z}{L}\right) \tag{14}$$

and

$$\frac{\partial S}{\partial z} = \frac{S_*}{kz}\phi_h\left(\frac{z}{L}\right) \tag{15}$$

where the stability functions ϕ_m and ϕ_h have been obtained empirically from carefully designed field studies. For a neutral PBL, $\phi = 1$; for an unstable PBL, $\phi < 1$; and for a stable PBL, $\phi > 1$. These expressions can be integrated as in (6), (7), and

(8) to relate the fluxes to measurements of velocity and scalar differences at two heights in the surface layer.

A similar procedure is used in the mixed layer, with the scaling height being the depth of the CBL, z_i and the velocity scale being the *Deardorff velocity*,

$$w_* = (F_{b0} z_i)^{1/3} \tag{16}$$

However, in the mixed layer a further complication is that the behavior of mixed-layer variables depends not only on surface fluxes but also on fluxes through the top of the CBL, the *entrainment fluxes*. Therefore, for scalar fluxes in the mixed layer both the surface flux F_{s0} and the entrainment flux F_{szi} need to be incorporated in generalized formulations. For the scalar flux-gradient relationship, this can be expressed as

$$\frac{\partial S}{\partial z} = -\frac{F_{s0}}{w_* z_i} g_0\left(\frac{z}{z_i}\right) - \frac{F_{szi}}{w_* z_i} g_{zi}\left(\frac{z}{z_i}\right) \tag{17}$$

where $g_0(z/z_i)$ and $g_{zi}(z/z_i)$ are the normalized mixed-layer gradient functions. Thus far, these gradient functions have not been measured in the atmosphere; however, they have been estimated from detailed numerical simulations of the CBL.

Usually it is the density of the trace constituent that is measured since most sensors respond to the number of molecules in a particular volume of air. In estimating the flux of a species, we normally calculate the quantity $\overline{ws}$ with the assumption that $W = 0$ at the surface. This is not strictly true even over a horizontally homogeneous surface if the water vapor and temperature fluxes are not zero. This arises from the constraint that the flux that is most realistically zero at the surface is the mass flux of dry air, $\overline{\rho w} = 0$. Intuitively, we can see that in the case of a heated surface, rising parcels of air will be on average warmer and lighter, and consequently contain fewer molecules per unit volume than their surroundings, while descending parcels will be colder and denser, and contain more molecules than their surroundings so that for zero species flux at the surface, $\overline{ws} < 0$. This is known as the *Webb effect*. To obtain the correct flux, it is necessary to correct for $W \neq 0$ by incorporating terms proportional to the fluxes of humidity and temperature. This correction becomes significant if $\overline{ws}/S$ is less than about 0.01 m/s. Alternatively, if instead of measuring the constituent density, we measure its mixing ratio with respect to dry air, there is no Webb correction. In subsequent discussion we disregard this correction, but note that it can be important for surface fluxes of relatively long-lived atmospheric species such as CO_2, CH_4, or N_2O.

5 OBSERVATIONAL TECHNIQUES

Since the boundary layer is a conduit for transport of trace species between the surface and the overlying free troposphere, measuring species fluxes within the PBL is a standard approach for estimating their sources or sinks at the surface, as well as

their rates of exchange with the overlying atmosphere. There are many ways to measure fluxes in the PBL. However, the two most widely used platforms are: (1) tower measurements in the surface layer and (2) airplane measurements in the mixed layer. There are, of course, other platforms that are used. For example, in the marine surface layer, ship-mounted instruments are used and in the mixed layer tethered balloons and neutrally buoyant airships have been used. The most direct and fundamental flux measurement technique is the eddy correlation technique [Eq. (1)]. However, this requires fast-response high-resolution measurements of species concentration and vertical air velocity over a time period or distance long enough to obtain a sufficiently accurate average of the turbulent fluctuations. As a rule of thumb, to estimate the flux to about 10% accuracy, one should average over several times the maximum eddy size making significant contributions to the flux. Generally, measuring fluxes from a tower a few meters above the surface for moderate winds requires an averaging time of about 20 min, which is about the same as that required for measuring fluxes from an aircraft in the middle of the mixed layer flying at 100 m/s.

At the other end of the spectrum, the smallest scale eddies that need to be measured to estimate a flux in the surface layer are roughly about $0.5z$. In the mixed layer, the smallest scales are roughly about $0.1z_i$. Therefore, for measurements from a tower at a height of 2 m, and a wind speed of 5 m/s, a frequency response of least 5 Hz is required. In the mixed layer, for an aircraft flying at 100 m/s in the middle of a 1-km-deep CBL, a frequency response of at least 1 Hz is required. If the aircraft is flying lower, say about 30 m, which is in the upper part of the surface layer, a frequency response of at least 7 Hz is required. [To achieve a frequency response of f_c hertz using a sensor with a first-order time response, a sensor time constant of about $1/(6f_c)$ s is required.]

In carrying out flux measurements by eddy correlation, both vertical velocity and species concentrations must be measured concurrently. *Sonic anemometers* are often used for the vertical velocity measurement from towers since they have good velocity resolution, adequate time response, and no moving parts. The air velocity component along the path between two sets of sonic transducers is obtained from the difference in the velocity of sound traveling along the same path in opposite directions. In addition, since the speed of sound is approximately proportional to the square root of virtual temperature, sonic anemometers are usually configured to also measure the virtual temperature, and thus the buoyancy flux can be obtained as well. Three-axis sonic anemometers are available commercially for measuring eddy correlation fluxes in the surface layer.

Measurement of air velocity components from aircraft requires measuring both the velocity of the air with respect to the aircraft and the velocity and angular orientation of the aircraft with respect to Earth. The former is often obtained from pressure measurements on the nose of the aircraft or from a probe mounted on a noseboom ahead of the aircraft. Pressure difference measurements are sensed from sets of ports. A forward-looking and a static pressure port are used to sense the airspeed, and sets of ports at different angles in both the horizontal and vertical plane of the aircraft are used to sense the flow angles of the air. The aircraft orientation and

velocity are often obtained from an inertial navigation system (INS) which senses the attitude angles and acceleration of the aircraft. The acceleration components are then integrated to obtain the velocity, and integrated again to obtain the position of the airplane. Often, navigational information from the satellite-based Global Positioning System (GPS) is used to remove drift inherent in the INS due to integration of a bias in the accelerometers. The air velocity is obtained from the difference between the velocity of the air with respect to the airplane, which is rotated by means of the attitude angles to an Earth-based coordinate system, and the velocity of the airplane, which is also measured in an Earth-based coordinate system.

Several techniques have been used to measure species concentration with sufficient resolution and frequency response that direct eddy correlation fluxes can be obtained. Water vapor fluxes have been obtained from both infrared and ultraviolet absorption devices. Fluxes of several other trace gases can also be sensed by infrared absorption, including CO_2, CH_4, and CO. Chemiluminescence is another inherently fast technique useful for ozone, isoprene, and possibly NO, NO_2, and dimethyl sulfide. In this technique, a reactive gas is mixed with the air, which reacts with the species being measured, with the resulting emission of photons detected by a photomultiplier tube. Finally, a tandem mass spectrometer, which ionizes, accelerates, and segregates the target species molecules has been used for measuring fluxes of acetone, ammonia, and formic acid in the surface layer.

Nearly all the techniques listed above (except for some open-path radiation absorption devices) require the air to be drawn into a sensing chamber of some sort. This requires careful consideration of the ducting system to ensure that the flow is fast enough and the ducting short enough that significant attenuation of concentration fluctuations does not occur in the frequency region with significant contributions to the flux. Generally this means that if the duct is longer than a couple of meters, the Reynolds number of the flow in the duct,

$$\mathrm{Re} = \frac{dU_t}{\nu} \tag{18}$$

where d is the tube diameter, U_t the flow velocity in the tube, and ν is the kinematic molecular viscosity ($\simeq 0.15 \times 10^{-4}\ \mathrm{m^2/s}$ for air at room temperature), must be greater than the critical value for turbulence to exist in the tube; i.e., $\mathrm{Re} > 2300$.

In addition to direct eddy correlation, several other techniques have been used for flux measurement. Most of these alternatives are implemented to relax the high-frequency requirements of direct eddy correlation. Conceptually, perhaps the simplest approach is to make measurements of species concentration less frequently, but grab the sample quickly so as to still retain the required frequency response. By this disjunct sampling technique, a flux can be estimated even if the frequency response of the concentration measurement is reduced by nearly an order of magnitude below what is required for direct eddy correlation. Another approach, called *eddy accumulation*, is to collect the air sample at a rate proportional to the vertical velocity, with the upward-moving air going into one reservoir and downward-moving air into another. The flux is then proportional to the difference in concen-

tration between the two reservoirs. With this approach, there is no longer any requirement for fast-response species measurement. In effect, the requirement for fast response is shifted to the flow control. Disadvantages of this approach are the small concentration difference between the two reservoirs and the requirement for fast-response and accurate flow control.

There are many other techniques for estimating flux, mostly with the objective of reducing the high-frequency response requirement, but, in contrast to the above, these approaches utilize some empirical relationship between the flux and some other variables. One simplification of eddy accumulation, called *relaxed eddy accumulation*, is to collect the air at a constant rate, regardless of the magnitude of the vertical velocity, in either of the two reservoirs depending on the sign of the vertical velocity. The flux then depends, in addition to the concentration difference between the two reservoirs, on the standard deviation of the vertical velocity and a parameter that depends on the vertical velocity distribution.

Measuring the gradient of species concentration either in the surface layer or the mixed layer is also used to estimate the surface flux. In the surface layer, the flux can be estimated from the integral of Eq. (15); i.e., from a difference in concentration between two levels plus the friction velocity, u_*, and a measure of the stability L, which depends on u_* and F_{b0}. Again, this does not require fast-response concentration measurements, but it does require measurement of small differences in concentration, as well as estimates of buoyancy and momentum fluxes.

In the mixed layer, Eq. (17) can similarly be integrated and solved for both the surface and the entrainment fluxes from mean concentration differences. However, since there are now two unknowns, mean concentration must be measured at a minimum of three levels to obtain two concentration differences unless one of the fluxes is estimated by another technique. Typical values of the normalized gradient functions have been estimated from large-eddy numerical simulations of the CBL to be, for $g_0(z/z_i)$ about 13 at $z/z_i = 0.1$ and about 1 at $z/z_i = 0.5$, and for $g_{zi}(z/z_i)$ about 70 at $z/z_i = 0.9$ and about 3 at $z/z_i = 0.5$. Since a typical value for w_* is about 1 m/s, we see that by taking the ratio of (17) to (5) the mixed-layer gradient is roughly about 1% of the surface layer gradient. This is again a reflection of the relative efficiency of transport in the mixed layer compared to the surface layer. This relatively small mixed-layer gradient is offset to a considerable extent by the much larger height differences that can be used in the mixed layer. Nevertheless, the concentration differences obtained by integration of the surface layer gradient formulation (5) can be several times larger than the differences obtained from integration of the mixed-layer gradient formulation (17). Thus far, the mixed-layer gradient technique has been used to estimate surface fluxes of isoprene and dimethyl sulfide, both of which have sources only at the surface and lifetimes of less than a couple of days, which reduces their concentration above the CBL to near zero.

Both surface layer and mixed-layer similarity relationships have also been obtained for scalar variance profiles. These relationships are based on the hypothesis that CBL variance is generated solely by surface and entrainment fluxes. In practice, this may have advantages for measuring flux, particularly in the mixed layer if fast-response scalar measurements are practicable but concurrent vertical velocity

measurements are not, since the mean concentration differences can be small. On the other hand, in practice mesoscale variability may contribute to the measured scalar variance and may be hard to estimate or remove from the measured variance.

Other less direct techniques exist for measuring constituent fluxes. One approach is to assume that the transport characteristics of a tracer species in the surface layer are the same as the species under consideration. Then if both eddy correlation fluxes and concentration differences are available for the tracer species, and only difference measurements are available for the species under consideration, the ratio of the unknown flux to the known flux is equal to the ratio of the tracer species difference to the unknown species difference.

Another approach is to use the budget equation of the species to solve for the surface (or entrainment) flux. The budget equation for the mean concentration of a species is given by

$$\frac{\partial S}{\partial t} + U(z)\frac{\partial S}{\partial x} + \frac{\partial F_s}{\partial z} = Q_s \tag{19}$$

where Q_s is the internal (e.g., chemical) source or sink of S, and we have assumed, for simplicity, that $V = W = 0$. This can be integrated, e.g., from the surface up to a height z and solved for the surface flux to obtain

$$F_{s0} = \frac{\partial \langle S \rangle}{\partial t} + z\langle U \rangle \frac{\partial S}{\partial x} + (\overline{ws})_z - z\langle Q_s \rangle \tag{20}$$

where $\langle\ \rangle$ denotes an average over the layer from the surface to height z. This approach has been used by aircraft flying in a Lagrangian flight pattern—i.e., advecting the flight pattern with the PBL mean wind using constant-level balloons as tracers, so that the second term on the right side of (20) is zero—and carrying out a series of flights over a day or more. In this case, the surface flux is obtained from the residual of the time rate of change, the entrainment flux, and the chemical source/sink terms.

REFERENCES

Fowler, D., and J. H. Duyzer, Micrometeorological techniques for the measurement of trace gas exchange, in M. O. Andreae and D. S. Schimel (Eds.), *Exchange of Trace Gases between Terrestrial Ecosystems and the Atmosphere*, Wiley, New York, 1989, 189–207.

Garratt, J. R. *The Atmospheric Boundary Layer*, Cambridge University Press, New York, 1992.

Kaimal, J. C., and J. J. Finnigan, *Atmospheric Boundary Layer Flows: Their Structure and Measurement*, Oxford University Press, New York, 1994.

Lenschow, D. H., Aircraft measurements in the boundary layer, in D. H. Lenschow (Ed.), *Probing the Atmospheric Boundary Layer*, American Meteorological Society, Boston, MA, 1986, pp. 29–55.

Lenschow, D. H., and B. B. Hicks (Eds.), *Global Tropospheric Chemistry: Chemical Fluxes in the Global Atmosphere*, National Center for Atmospheric Research, Boulder, CO, 1989.

Matson, P. A., and R. C. Harriss, *Biogenic Trace Gases: Measuring Emissions from Soil and Water*, Blackwell Science, Cambridge, MA, 1995.

Raupach, M. R., Canopy transport processes, in W. L. Steffen and O. T. Denmead (Eds.), *Flow and Transport in the Natural Environment: Advances and Applications*, Springer-Verlag, Berlin, 1988, 98–127.

Stull, R. B., *An Introduction to Boundary Layer Meteorology*, Kluwer Academic, Dordrecht, The Netherlands, 1988.

Wesely, M. L., Turbulent transport of ozone to surfaces common in the eastern half of the United States, in S. E. Schwartz (Ed.), *Trace Atmospheric Constituents: Properties, Transformations, and Fates*, Wiley, New York, 1983, pp. 346–370.

Wyngaard, J. C., On the maintenance and measurement of scalar fluxes, in T. J. Schmugge and J.-C. André (Eds.), *Land Surface Evaporation: Measurement and Parameterization*, Springer-Verlag, New York, 1991, pp. 199–229.

CHAPTER 11

SOURCES AND COMPOSITION OF AEROSOL PARTICLES

RICHARD ARIMOTO

1 INTRODUCTION

The atmospheric aerosol is a suspension of solid and liquid particles in the air that displays a degree of stability with respect to gravitational settling. Aerosol particles originate from a large number of sources whose influences can change dramatically over time scales of minutes to hours or can remain relatively constant for years. Although the term *aerosol* technically applies to both the solid and liquid particles and the gases in which they are suspended, common usage allows *aerosol* to refer to the particles alone, a practice that will be followed here. Increasing interest in the sources and composition of aerosols has resulted from a growing awareness of their linkages to meteorology, climate, and global change and from a better appreciation of the roles these particles play in biogeochemical cycles.

Aerosols range in size from clusters of molecules (≤ 0.001 μm radius) to ultra-giant particles with radii of 100 μm or more. In one commonly used scheme (Junge, 1963), the particle size spectrum for the aerosol is separated into three classes: (1) Aitken particles (≤ 0.1 μm radius); (2) large particles (0.1 to 1.0 μm radius); and giant particles (> 1 μm radius). In another scheme (Whitby, 1978) particles with radii, $r < 0.1$ μm, which are formed by homogeneous condensation, are referred to as the nucleation mode; particles in the 0.1 to 1-μm size range are referred to as the accumulation mode because they are formed from the accumulation of nucleation mode particles and the deposition of gases. Often aerosols < 1 μm radius are simply referred to as fine particles (e.g., Heintzenberg, 1989), while larger aerosols, with a peak in the mass distribution at $r = 2$ to 5 μm, compose the coarse mode.

The various schemes for characterizing aerosol size distributions generally rely on the concept of an aerodynamic equivalent size, which is a normalization based on

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

the behavior of a spherical particle of unit density (1 g/cm^3). Most solid aerosol particles, however, are not spherical and few are of unit density. Some aerosol particles, such as sea salt under low relative humidity, are crystalline while others, including many composed of mineral matter, are angular. Aerosol particles also take the shape of rods or flat plates, and both natural and anthropogenic aerosols can be aggregates of smaller particles, which can be roughly spherical or in some cases can form chains. Biological aerosols, especially spores and pollen, often display complex geometries that have evolved to favor dispersal over long distances.

Populations of aerosol particles can be classified according to various criteria in addition to size: natural versus anthropogenic, organic versus inorganic, internally mixed versus externally mixed, mechanically generated (primary particles) versus products of gaseous reactions (secondary), etc. These various classification schemes often serve specific purposes, and the diversity in the different types of aerosols can hardly be overemphasized. This chapter will serve as an introduction to the sources and composition of the particles that compose the atmospheric aerosol. It will also summarize information regarding the sources and composition of aerosols and introduce some of the ways in which the biogeochemical cycles of aerosols are linked to those of other atmospheric constituents.

2 MECHANICALLY GENERATED AEROSOLS

Particles produced by mechanical processes tend to be larger than those resulting from gas-to-particle conversion. In general particles larger than a micrometer are mechanically formed by processes such as the wind erosion of soils, the bursting of bubbles in seawater, the shedding of plant fragments, etc. Although the relationship between the size of an aerosol particle and length of time it remains suspended in the atmosphere is complex, larger particles generally fall out of suspension more quickly than smaller ones (Fig. 1); hence large mechanically generated particles tend to have comparatively short atmospheric residence times.

Even though most mechanically generated aerosols are removed from the atmosphere close to their sources, some coarse particles remain suspended in the atmosphere for weeks and can travel thousands of kilometers before finally being deposited. While they are in suspension, aerosol particles can react with gases, with hydrometeors, and with other particles. As illustrated below, such reactions link the cycles of various atmospheric constituents in complicated ways.

The strengths of the various aerosol sources can be evaluated in several ways, and one of the most straightforward is to consider the mass of material injected into the atmosphere. As mechanical sources tend to produce physically and aerodynamically large particles, the importance of these sources is most evident when mass fluxes or related characteristics, such as particle volume, are being considered. In contrast, when evaluating source strengths with respect to the numbers of particles produced, the contributions from the mechanical sources tend to be less important compared with those producing numerous small particles via gas-to-particle conversion.

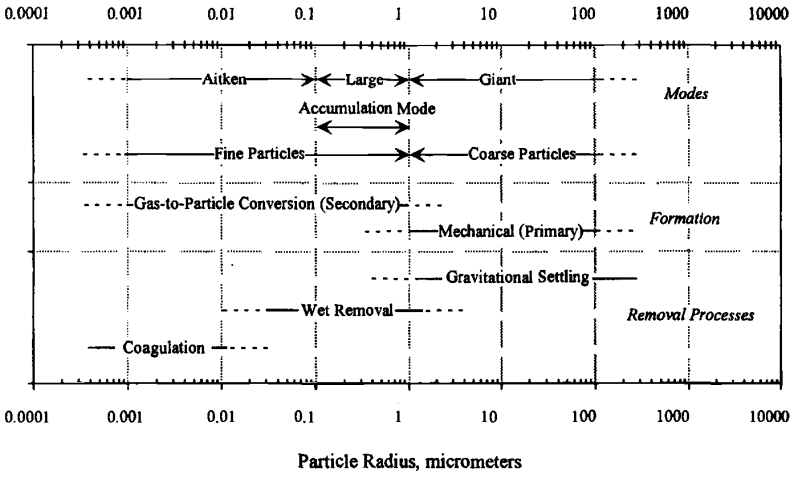


Figure 1 Characteristics of aerosol particles and the processes by which they are removed from the atmosphere.

Mineral Aerosol

The physical and chemical weathering of Earth's continental crust results in the production of mineral aerosol particles, commonly called atmospheric dust; and this represents one of the largest sources on a mass basis for natural particulate material in the atmosphere. Chinese records of dust storms date back thousands of years, and plumes of mineral aerosol over the oceans have been observed by mariners since humans have taken to the sea. Modern technology has shown that dust plumes over the oceans are among the most dramatic features seen in satellite images of aerosol optical depth (Husar et al., 1997).

Worldwide, about a third of Earth's surface can be considered potential sources for dust, but the arid and semiarid lands in Africa and Asia are the largest sources (Fig. 2). Climate clearly affects the amount of atmospheric dust produced. In general more dust is generated as the land becomes drier, but in hyperarid areas deserts can become "blown out" and less important as sources. Drought cycles also are linked to the emissions of desert dust. For example, studies at Barbados, an island in the North Atlantic Ocean, have shown that atmospheric dust concentrations increased during the Sahelian drought of the early 1970s (Prospero and Nees, 1977). Dust loads in the atmosphere also can vary over longer periods of time as a consequence of large-scale changes in climate and circulation. In this context, studies by An et al. (1990) suggest that patterns in dust deposition to the Chinese loess plateau over thousands of years can be linked to variations in the strength of the Asian winter monsoon.

The exact amount of dust injected into the atmosphere remains uncertain, but recent estimates are of the order of ~1500 Tg/yr (Andreae, 1995; Tegen et al.,

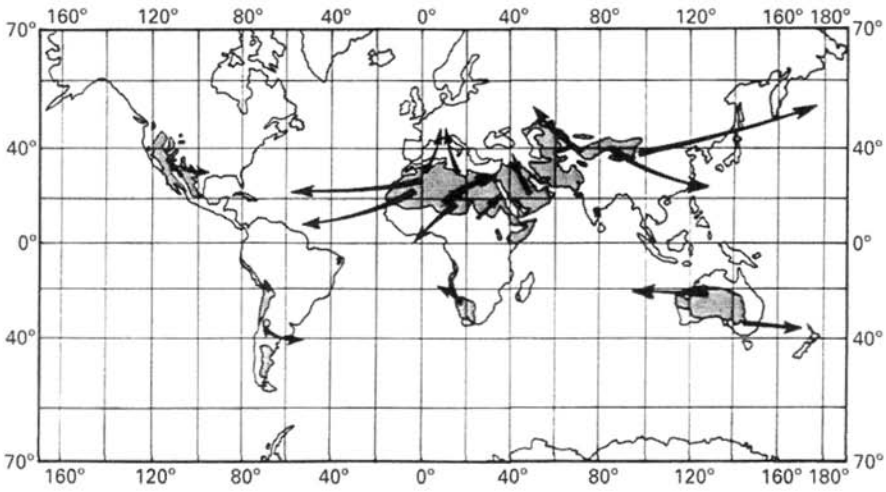


Figure 2 Sources for mineral aerosol (atmospheric dust). (From Péwe, 1981.)

1996), with an uncertainty of perhaps a factor of 2. As human activities have altered the global landscape, a portion of the atmospheric dust load can be considered anthropogenic. For example, modeling studies by Tegen et al. (1996) indicate that $50\% \pm 20\%$ of the global dust flux may come from disturbed soils. On the other hand, efforts made by humans to reclaim some desert lands (Parungo et al., 1994) may have reduced the strength of natural dust sources.

Mineral particles are formed by a variety of processes, including grinding, weathering, abrasion, etc. (Pye, 1987). Once the particles are formed, the wind deflates and disperses them, but other factors, such as the sizes and shapes of the particles, the roughness of the particle bed, the cohesiveness of the particles, the presence of cementing agents, the extent of vegetative cover, and especially the amount of soil moisture influence the erodibility of the soils. Studies of the dynamics of the dust generation process by Gillette et al., (1974) showed that sandblasting of the soils was the dominant mechanism for mineral aerosol production by wind erosion, and these and other authors have shown that the wind velocity also shapes the size distributions of the suspended dust particles.

The transport of desert dust affects the global cycles of nitrogen, phosphorus, sulfur and various trace elements (Prospero, 1981; Schlesinger et al., 1990). Some areas of Earth benefit from the transport and deposition of mineral dust; for example, the fertility of the Loess Plateau in central China results in large measure from the accumulation of nutrient-rich mineral particles transported through the atmosphere from deserts in northwestern China (Liu et al., 1985). Similarly, dust originating from the Sahara Desert is transported through the atmosphere to the Central Amazon Basin where it supplies critical trace elements (Swap et al., 1992). Other parts of the continents are stripped of nutrients by the combined actions of wind and water erosion, and the economic consequences of erosion are substantial. For example,

cost estimates for lost agricultural productivity, damage to waterways and infrastructure, and public health problems due to erosion by wind and water run into the billions of dollars for the United States alone (Pimentel et al., 1995).

The transport and deposition of mineral aerosol affects the cycles of a large number of trace elements in addition to those of N, P, and S. Bulk atmospheric dust particles generally have an elemental composition similar to that of average crustal rock (Rahn, 1976), and the composition of the ambient aerosol often is evaluated through "enrichment factors" (EFs) which are defined as

$$\mathrm{EF(Al, Crust)} = \frac{(\mathrm{X/Al})_{\mathrm{Aerosol}}}{(\mathrm{X/Al})_{\mathrm{Crust}}}$$

where X is any element of interest; Al is aluminum, a commonly used reference element; and the subscripts Aerosol and Crust refer to the aerosol sample of interest and the crustal reference material, respectively. Another commonly used reference element is Si, but Sc or a variety of other elements would serve the purpose equally well.

Weathered crustal material is the presumptive source for any element whose enrichment factor for a given sample approaches unity; those elements with EFs greater than ~5 have significant noncrustal sources. Direct comparisons of elemental ratios in aerosol samples versus crustal rock also show that the atmospheric loadings of mineral dust govern the concentrations of a large number of trace elements in the atmosphere (Table 1). It is important to point out, however, that individual mineral dust particles can have a composition quite different from either the bulk dust or average crustal material (Anderson et al., 1996).

In a study of erodible soils, Schütz and Rahn (1982) showed the concentrations of most elements increased as particle sizes decreased to 20 to 50 μm radius, but the concentrations reached a plateau for particles less than 10 to 20 μm in radius. These authors predicted that some variability in the elemental composition of dust should occur near the desert source areas where a significant fraction of the particles would have radii > 10 μm. More than ~1000 km from the sources, however, the bulk of the particles would be < 10 μm in radius, and therefore these authors concluded that the elemental composition of dust transported long distances would be similar to that of the continental crust.

Dust particles in the atmosphere are far from inert, and reactions occurring on dust particles have significant implications for several important chemical cycles. Direct observations of individual particles showed sulfate coatings were present on >40% of the mineral dust particles collected over the North Atlantic at 25°N, and nitrate coatings were observed on >30% of the particles (Parungo et al., 1986). Further evidence for the uptake of gaseous sulfur species on dust from the Asia–Pacific region was obtained through statistical analyses of the elemental composition of aerosols (Winchester and Wang, 1990). Reactions between dust particles and gaseous nitrogen oxides have been reported from laboratory studies (Mamane and Gottlieb, 1992) and from analyses of ambient aerosols (Wu and Okada, 1994). The formation of nitrate on dust particles via heterogeneous reactions constitutes a sink

TABLE 1 Mass Ratios of Crustal Elements to Aluminum for High-Dust Events at Barbados, Bermuda, and Izaña

	Observed			
Element	Barbados	Bermuda	Izaña	Average Crustal Rock[a]
Ba	6.2×10^{-3}	1.2×10^{-2}	9.8×10^{-3}	6.8×10^{-3}
Ca	2.9×10^{-1}	3.3×10^{-1}	3.6×10^{-1}	3.7×10^{-1}
Co	2.4×10^{-4}	2.4×10^{-4}	3.0×10^{-4}	1.2×10^{-4}
Cr	1.1×10^{-3}	1.9×10^{-3}	1.5×10^{-3}	4.4×10^{-4}
Cs	4.7×10^{-5}	6.2×10^{-5}	7.1×10^{-5}	4.6×10^{-5}
Eu	2.2×10^{-5}	2.3×10^{-5}	2.9×10^{-5}	1.1×10^{-5}
Fe	5.1×10^{-1}	6.1×10^{-1}	7.0×10^{-1}	4.4×10^{-1}
Hf	5.2×10^{-5}	5.8×10^{-5}	8.1×10^{-5}	7.2×10^{-5}
Mg	3.7×10^{-1}	3.2×10^{-1}	3.0×10^{-1}	1.6×10^{-1}
Mn	1.1×10^{-2}	9.5×10^{-3}	1.2×10^{-2}	7.5×10^{-3}
Na	1.1×10^{0}	3.3×10^{-1}	1.1×10^{-1}	3.6×10^{-1}
Rb	1.1×10^{-3}	1.7×10^{-3}	1.6×10^{-3}	1.4×10^{-3}
Sb	1.2×10^{-5}	1.3×10^{-5}	1.4×10^{-5}	2.5×10^{-6}
Sc	1.7×10^{-4}	2.0×10^{-4}	2.3×10^{-4}	1.4×10^{-4}
Ta	2.1×10^{-5}	2.1×10^{-5}	2.8×10^{-5}	2.7×10^{-5}
Tb	1.6×10^{-5}	1.5×10^{-5}	1.9×10^{-5}	8.0×10^{-6}
Th	1.6×10^{-4}	2.0×10^{-4}	2.0×10^{-4}	1.3×10^{-4}
V	1.5×10^{-3}	3.5×10^{-3}	1.6×10^{-3}	7.5×10^{-4}
Yb	4.2×10^{-5}	5.0×10^{-5}	5.2×10^{-5}	2.7×10^{-5}

[a]Taylor and McLennan (1985).

for nitrogen oxides, and reactions involving dust, N_2O_5, O_3, and HO_2 radicals may affect the cycles of photochemical oxidants, leading to decreases in tropospheric ozone near dust sources (Dentener et al., 1996).

Mineralogical studies have shown that atmospheric dust consists of silicates (quartz and feldspars); clay minerals (e.g., kaolinite, smectite, illite, mica), carbonates (calcite and dolomite), and sulfur minerals (gypsum and anhydrite) (see Pye, 1987). Mineralogical analyses of aerosol particles by Zhou and Tazaki (1996) have provided independent lines of evidence for the chemical reactivity of atmospheric dust. Their analyses showed S-rich submicrometer particles frequently are found attached to mineral dust particles, and they inferred that H_2SO_4 reacted with calcite during transport to form gypsum.

The selective removal of dust particles as a function of particle size during transport probably has little effect on elemental composition, except perhaps for the rare earths (Sholkovitz et al., 1993), but size fractionation can lead to mineralogical differences among samples (Johnson, 1976). Glaccum and Prospero (1980) similarly suggested that the proportion of quartz particles relative to clay minerals should be low in dusts that have traveled long distances owing to the preferential fallout of the quartz particles, which tend to be aerodynamically large. Even so giant

quartz particles ~50 μm radius have been found in the atmosphere over the central North Pacific, thousands of kilometers from their sources (Betzer et al., 1988). An unresolved paradox confronting atmospheric scientists is that the presence of such large particles so far from their sources is difficult, if not impossible, to explain based on our current understanding of transport dynamics.

3 SOURCES PRODUCING PRIMARY AND SECONDARY PARTICLES

Primary Particles from Oceans: Sea Salt Aerosol

Surface winds cause the production of aerosols from the sea as well as from the land. The effects of the wind over the ocean are mediated by breaking waves, bursting bubbles, and to a lesser extent the formation of large spume droplets torn from waves by strong winds. The seasalt injected into the atmosphere is another of the large sources for aerosols on a mass basis, of the order of 1000 to 10,000 Tg/yr (Blanchard, 1983). One reason for the large uncertainty in this estimate is that there is no strict definition of what constitutes a sea salt aerosol particle, i.e., very large particles have such short atmospheric residence times that one might question whether they are truly suspended in the atmosphere.

The production of sea salt particles is mainly due to oceanic whitecaps. Bubbles from the whitecaps burst at the sea surface producing film droplets and jet droplets, which have quite different properties and whose proportions vary as a function of bubble size (Blanchard, 1980). Model estimates by Erickson and Duce (1988) indicate the mass median radii (MMR) for sea salt over the oceans (50% of the sea salt mass occurs on particles smaller than the MMR and 50% on particles larger than the MMR) should range between 3.0 and 7.5 m, which is in good agreement with observations. Both the amount of salt produced and the sizes of the particles vary in response to wind speed, but a consideration of the relative amounts of ocean and land covering Earth's surface together with the atmospheric loadings of dust and sea salt shows the production of sea salt particles may be less efficient than dust.

For many substances, including sulfate, the composition of fresh, bulk, sea salt aerosols is similar to that of seawater, but as noted for mineral aerosol, one must recognize that the elemental composition of individual sea salt particles may be quite different from that of the bulk aerosol. Once the jet and film drops are ejected into the atmosphere, the water in them begins to evaporate, leading to a droplet of high ionic strength, which can undergo repeated cycles of dilution and concentration. Fractionary recrystallization within the evaporating drops can be followed by shattering of the particles, and this process can lead to variations in the content of elements such as Mg, S, K, and Ca among individual sea salt particles (Mouri et al., 1993).

Some substances, including some heavy metals, organic matter, radionuclides, and nutrient species are enriched in the sea salt aerosols as a result of the scavenging of surface-active material as bubbles pass through the water column and rupture the sea surface microlayer (e.g., MacIntyre 1974; Wallace and Duce, 1975). The enrich-

ments of trace elements in experimentally produced sea salt particles can reach several tens of thousands (Weisel et al., 1983), and this enrichment process can lead to a recycling of material between the surface ocean and the atmospheric marine boundary layer.

Sea-salt particles contain variable amounts of organic material, and large salt particles from over the remote oceans typically contain organic carbon with an isotopic composition similar to that of source materials in seawater (Buat-Ménard et al., 1989). The carbon concentrations in sea salt particles (normalized to Na) are several 100-fold higher than that of seawater, presumably as a result of the same physicochemical processes causing the enrichments of inorganic materials (Wallace and Duce, 1975). In some areas of the oceans, terrestrial sources also can contribute significant amounts of carbon to large particles, but those continental sources, whether anthropogenic or natural, are more important for submicrometer marine aerosols.

Although the organic composition of marine aerosols is only partially characterized at best, numerous organic compounds have been detected in marine aerosols. For example, Peltzer and Gagosian (1989) investigated aliphatic hydrocarbons, wax esters, fatty alcohols, sterols, fatty acids, and long-chain unsaturated ketones in aerosols from several sites in the Pacific Ocean. These compounds were used as biomarkers in studies of the sources, transport, and transformation of organic material in the marine atmosphere. Kawamura and Usukura (1993) investigated dicarboxylic acids in the western North Pacific, and they concluded the diacids were mainly from Asia, but some diacids were produced by photochemical reactions in situ.

Sea-salt particles also react with a variety of gaseous components of the marine atmosphere; most notably HNO_3, methanesulfonic acid, and H_2SO_4 are sorbed by liquid sea salt particles and HCl and HF are displaced (Ericksson, 1960; Okada et al., 1978). The modeling of acid displacement reactions is made difficult by the nonideal behavior of the high solute concentrations in the sea salt droplets (Brimblecombe and Clegg, 1988). However, analyses of individual particles from the North Atlantic suggest that Cl loss from sea salt can be accompanied by the formation of $NaNO_3$ (Pósfai et al., 1995). Reactions of sea salt with various N gases, such as NO_2, $ClNO_3$, and N_2O_5, have been observed, and those reactions could lead to the formation of $NaNO_3$ (Schroeder and Urone, 1974; Finlayson-Pitts et al., 1989; Keene et al., 1990). Such reactions also could generate reactive Cl atoms; and analogous to the hydroxyl radical, the atomic chlorine would participate in photochemical reactions with various organic substances.

Other aqueous-phase reactions in sea salt aerosols involve the oxidation of SO_2 by O_3 to non-sea-salt (NSS) sulfate (i.e., the sulfate in excess of what can be attributed to sea salt from unfractionated seawater) (Sievering et al., 1992, 1995; Chameides and Stelson, 1992). Recent analyses by Keene et al. (1998) indicate that the oxidation of SO_2 by ozone in sea salt aerosols is only a minor source for NSS sulfate, and these authors suggested that oxidation of SO_2 primarily occurs via another pathway, possibly involving HOX, where X is chlorine or bromine. As SO_2 originates from anthropogenic as well as natural sources, these reactions not

only show how heterogeneous reactions can affect the composition of aerosols but also illustrate how intimately the chemistry of pollutants can be linked with the cycles of natural substances in the atmosphere.

Secondary Particles from Marine Sulfur Compounds

Some sources produce aerosols both by mechanical processes and by gas-to-particle conversion, the oceans being a good example of this. In addition to the mechanically generated sea salt aerosol, the oceans emit gaseous compounds that can be oxidized and eventually produce aerosol particles. The most important chemicals in this case contain sulfur, especially dimethylsulfide (DMS), methanesulfonic acid (MSA), and sulfate. Interest in the marine sulfur cycle increased enormously after a hypothesis was proposed connecting oceanic phytoplankton to aerosols to clouds and hence to climate (Charlson et al., 1987). In this hypothesis, DMS produced by marine phytoplankton evades from the ocean and is oxidized to sulfate aerosol (and MSA), which can act as cloud condensation nuclei (CCN). The number of CCN in the atmosphere affects the reflectivity of clouds (the cloud albedo), and in this way oceanic emissions of reduced sulfur gases are linked to climate. Sulfate aerosols, whether marine derived or originating from continental emissions, also can reflect solar radiation back to space, and in so doing influence weather and climate directly.

There is at present no universally accepted chemical mechanism for the formation of NSS sulfate via DMS oxidation. One of the controversies that arose with respect to the oxidation of DMS was whether sulfur dioxide (SO_2) was an important intermediate in the pathway leading to sulfate aerosol (Bandy et al., 1992; Lin and Chameides, 1993). Recent studies indicate that the pathway from DMS to NSS sulfate does indeed include SO_2 as an intermediate, but our knowledge of the other compounds and reactions in the pathway is far from complete (Keene et al., 1998). Whether from DMS or from pollution sources, SO_2 can be further oxidized both in the gas and liquid phase to H_2SO_4, but these processes are sensitive to gas-phase nitric acid and ammonia concentrations (Clegg and Toumi, 1997), again demonstrating links between the chemistry of aerosols and gaseous species.

Over vast areas of Earth's oceans, from about 30°N to 30°S, the ratio of biogenic NSS sulfate to MSA tends to be constant, so much so that a MSA/NSS sulfate mass ratio of ~18 to 20 has been used as a diagnostic for marine biogenic sulfate (Savoie and Prospero, 1989; Savoie et al., 1994; Arimoto et al., 1996). The relative amounts of MSA and NSS sulfate do change with latitude however; above ~30° (N or S) more of the biogenic sulfur occurs as MSA (Berresheim, 1987; Pszenny et al., 1989; Koga et al., 1991; Bates et al., 1992). One of the central issues of the marine sulfur cycle currently being investigated is the extent to which this latitudinal dependence in MSA/NSS sulfate ratios is driven by the temperature dependencies of various reactions in the DMS oxidation pathways versus the influences of other photochemically active compounds.

In the marine atmosphere gaseous H_2SO_4 either can deposit on existing surfaces, again potentially involving sea salt, or it can form new sulfate particles via homogeneous nucleation. A fundamental issue concerning the marine sulfur cycle has to

do with where new aerosol sulfate particles form. The cycling and fate of particles formed in the marine boundary layer (MBL) would be far different from those formed in the free troposphere, and therefore this issue has important implications for the direct and indirect effects of the aerosols on solar radiation. While new particles form in the marine boundary layer under certain conditions (Covert et al., 1992), most of the new particle production evidently occurs in the proximity of clouds (Hegg et al., 1990; Perry and Hobbs, 1994). Recent studies conducted for ACE-1 (aerosol characterization experiment) showed that few new particles formed in the MBL over the Southern Ocean south of Australia (Clarke et al., 1997). Instead layers of new particles, DMS, MSA, and H_2SO_4 were observed in the free troposphere in the outflow regions of clouds at altitudes of several kilometers. Andreae and Crutzen (1997) suggest the DMS–aerosol–climate connection may still pertain because the subsidence of aerosol-laden air from the free troposphere into the MBL can supply particles that are initially too small to act as CCN but through heterogeneous or cloud processes can grow and become CCN.

Volcanoes

Volcanoes are another natural source producing both primary and secondary aerosol particles. One of the distinctive aspects of volcanic emissions is that strong eruptions can inject materials directly into the stratosphere where aerosol-induced effects on the balance of solar radiation and ozone depletion, for example, can persist for years (McCormick et al., 1995). Much of the work on volcanic aerosols has, in fact, focused on the stratosphere, but that topic will not be covered in this chapter.

The amounts of material produced by volcanoes can be quite considerable when they are active, but volcanic eruptions are episodic, and most volcanoes exhibit periods of dormancy following the releases of gases, particles, and lava that constitute the active phase. Explosive volcanoes eject primary particles, mainly silicate dust particles and ash, into the atmosphere; but many of the primary particles are so large that they settle out quickly and close to their source. Andreae (1995) observed that during periods of extreme volcanic activity, as much as 10,000 Tg of dust could be produced per year. An annual flux of that magnitude would be larger than that from any other aerosol source, with the possible exception of sea salt. In less active times, the flux of primary particles from volcanoes, ~4 Tg/yr, would be almost negligible on a global scale. The long-term average production of primary particles from volcanoes has been estimated as 33 Tg/yr (Andreae, 1995).

Volcanoes also release water vapor, CO_2, SO_2, fluorine, and chlorine into the atmosphere (Lambert et al., 1988; Symonds et al., 1988) both from explosive events and during noneruptive activity. Secondary particles, mainly sulfuric acid droplets, can form from these gaseous emissions but, on a mass basis, the production of secondary particles during periods of high volcanic activity is much smaller than that of primary particles. For nonexplosive, basaltic volcanoes and fissures, the masses of primary and secondary particles produced are more nearly comparable, but the combined production of primary and secondary particles from these sources is considerably smaller than during more active periods. More important, the volca-

nic emissions of sulfur (9.3 to 11.8 Tg S per year in total) are equivalent to 10 to 30% of the total anthropogenic sulfur flux into the atmosphere; this amount is large enough to significantly affect the chemistry of sulfur in the atmosphere (Berresheim et al., 1995).

Volcanoes inject substantial quantities of trace elements into the atmosphere, and as plumes of volcanic material cool, gaseous species condense and attach to particles. As a result the composition of volcanic aerosols typically is quite different from the parent magma. Compared with crustal material, volcanic particles tend to be enriched with relatively volatile elements, including Zn, Cu, Au, Pb, As, Cd, Sb, and Se (Buat-Ménard, 1990). These enrichments vary not only among different volcanoes but also during different eruptive stages for a particular volcano. The elemental composition of the aerosol is particularly sensitive to the amount of nondegassed magmatic material brought to the surface by the volcanic activity. Along this same line, iridium enrichments were found during the eruption of Kilauea (Zoller et al., 1983) but not six other volcanoes, presumably reflecting the different types of magma involved in the eruptions.

Globally volcanoes supply ~50% of the ^{210}Po in the atmosphere (Lambert et al., 1982). This nuclide is the last radioactive daughter in the decay series of the naturally occurring radionuclide ^{238}U. In Antarctica, volcanoes are a particularly important source for volatile radionuclides because snow and ice cover minimizes the impact of many other sources (Polian and Lambert, 1979). These authors found that the $^{210}Po/SO_2$ ratios in the plume from Mt. Erebus were 30-fold higher than those for areas not affected by volcanoes. The recognition of a strong volcanic source for ^{210}Po enabled Lambert et al. (1988) to estimate trace element fluxes from volcanoes by scaling them relative to ^{210}Po. The uncertainties in these figures are estimated to be a factor of 3, but in general the volcanic inputs of trace elements probably are $< 20\%$ of their total atmospheric inputs (Nriagu, 1989). One exception to this is Bi whose volcanic source strength is perhaps 10 times higher than the inputs from either natural or anthropogenic sources (Lee et al., 1986; Lambert et al., 1988).

Biological Aerosols

The winds not only generate aerosols but also scatter preformed particles, including pollen grains, spores, bacteria, viruses, algae, fungi, nematodes, protozoa, and fragments of plant and animal tissues. The concentrations of certain kinds of biological aerosols are monitored for allergy sufferers through the familiar air quality indices of fungal spores and pollen. More generally, however, investigations of biological aerosols have been limited despite their relevance for studies of air quality, climate, chemical cycles, and so forth. Biological aerosols span a large range in size, from radii of < 0.1 μm for viruses to hundreds of micrometers for large pollen grains and spores. Evolution has shaped certain types of pollen and spores to favor their dispersal through the atmosphere, and thus even though they are geometrically quite large, such particles can be transported over long distances and to great

heights. For example, culturable fungi have been recovered from the atmosphere at altitudes between 57 and 77 km (Imshenetsky et al., 1978).

Fungi are among the most abundant of the viable biological aerosols (Duce et al., 1983), and their numbers vary strongly with location, season, meteorology, and diurnal cycle. Worldwide, fungi of the genus *Cladosporium* are the most abundant, and in temperate regions fungal spores from this genus are especially abundant in summer and early fall. Spores and hyphal fragments from other genera of fungi, including *Alternaria*, *Drechslera*, *Epicoccum*, *Aspergillis*, and *Penicillium*, are commonly collected in samples of particulate matter in air. Under some circumstances, such as crop harvesting or mowing, the numbers of airborne fungal spores from local sources can reach impressive numbers, up to 10^9 spores per cubic meter of air (Levetin, 1995). Many species of fungi contain substances called allergans that trigger allergic reactions in humans, and various types of fungi cause respiratory and opportunistic infections. Leathers (1981) reported that each year several hundred thousand persons are infected with airborne, disease-causing fungi in the United States alone. Among the pathogenic fungi, *Coccidiodes imitis*, which is endemic to the southwestern United States and causes "valley fever," presents particularly serious health and economic problems. Each year several hundred persons require hospitalization because of this fungus, which is spread via spores transported through the atmosphere from the desert regions to metropolitan areas.

Bacteria are patchily distributed in the atmosphere, and they are released from both natural and anthropogenic sources by various mechanical processes including wind abrasion, agricultural activities, etc. Typical concentrations of culturable bacteria range from 10 to 1000 colony-forming units per cubic meter of air, but the numbers of bacteria in the atmosphere can reach 10^9 per cubic meter under disturbed conditions (Muilenberg, 1995). Bubbles rising in the oceans scavenge bacteria and viruses from the water column (Blanchard, 1983; Baylor et al., 1977, respectively) in the same way organic carbon and trace elements are scavenged, and bacterial enrichments of several 100-fold have been observed in the aerosol relative to seawater. Airborne bacteria produced in the operation of wastewater treatment plants pose potential health hazards (Hickey and Reist, 1975), but the best known case of a health problem associated with biological aerosols is Legionnaires' disease caused by bacteria (*Legionella pneumophilia*) growing in air-conditioning cooling towers (Dondero et al., 1980).

Pollen grains contain genetic material from male seed plants, and one group of pollen-producing plants has been classified as anemophilous because they entrust pollination of the female flowers to the wind rather than insects (Muilenberg, 1995). Pollen is produced in flowers, and the amount of airborne pollen is governed by the life cycles of plants, which in turn are influenced by extrinsic factors such as temperature, the photocycle, and precipitation. The walls of many pollen grains are composed of a resistant outer layer made of a polymer called sporopollenin and an inner wall of cellulose. Allergans associated with pollen most commonly affect the upper respiratory tract as hay fever and related maladies, but pollen exposure also can lead to asthma. The pollen from anemophilous plants, which are more common in temperate areas and less so in the tropics, tends to be smaller

than from the entomophilous (insect-pollinated) plants. Studies of pollen grains in marine sediments have been used in paleoclimate reconstructions, particularly those involving paleowinds (e.g., Hoogheimstra, 1987).

Aerosols from Biomass Burning

The burning of living and dead vegetation (biomass burning) is widespread over Earth, and this is a globally significant source for aerosols and for a variety of radiatively active and chemically reactive trace gases. Most of the biomass burned is caused by human activities as opposed to natural fires, and this mainly occurs in the tropics, involving savannas more than forests (Hao and Liu, 1994). Some burning sources are persistent, but emissions from savanna burning vary biennially because the growth of the savanna vegetation occurs during the wet season, and afterwards as biomass dries, it is burned.

Aerosols produced by biomass burning are composed of some black carbon (soot) but mainly organic carbon with hydrogenated and oxygenated functional groups. The amount of black carbon produced by fires is highly variable, and this is determined by the type of fuel consumed and whether the fire is in the ignition phase, flaming, or smoldering. Moreover, the composition of the aerosols can change quite rapidly—over time scales of seconds to minutes—as the particles are advected away from the fire. The transformation of particles in a smoke plume can lead to internal mixtures, i.e., particles with cores of black carbon and low-volatility organic compounds become coated with outer layers of more volatile organics (Mazurek et al., 1996).

The amount of particulate matter put into the atmosphere by biomass burning is estimated to be ~90 Tg per year (37 Tg/yr of this is from savanna burning), and this amounts to more than 20% of the total suspended particulates from all anthropogenic sources (Andreae et al., 1996). The black carbon from biomass burning (60 Tg/yr) accounts for an even larger fraction, i.e., two-thirds of the global black carbon emissions from all anthropogenic sources. These authors estimate that the number of cloud condensation nuclei generated globally by biomass burning activities is 35×10^{27}, over 10% of which is from savanna fires.

The organic carbon composition of biomass burning aerosols is not fully characterized owing at least in part to the diversity of fuels burned in different geographical regions. However, studies of tropical biomass burning have shown the major organic components are straight-chain, aliphatic, and oxygenated compounds, triterpenoids from plant waxes, resins/gums, and biopolymers (Simoneit et al., 1996). The fatty acid composition of aerosols produced in laboratory and field burns (Ballentine et al., 1996) was found to be dominated by saturated even-chain compounds, reflecting the importance of terrigenous plant waxes.

The polycyclic aromatic hydrocarbons (PAHs) in aerosols from biomass burning include biphenyl, trimethylnapthalenes, phenanthrene, anthracene, methylphenanthrenes, fluoranthene, pyrene, methylpyrenes, chrysene, benzanthracene, benzofluoranthenes, benzo ([e] and [a]) pyrenes, indenopyrene, benzo(ghi)perylene, and coronene (Simoneit et al., 1996; Ballentine et al., 1996). Simoneit et al. (1996)

also reported the occurrence of oxy-PAHs in burning products, including fluorenone, anthra-9,10-quinone, cyclopenta(def)phenanthrene-4-one, benzo[a]fluorene-11-one, benzanthrone, and napthanthrone. Both the PAHs and the oxy-PAHs are produced by incomplete combustion, and the production and transport of PAHs is of particular concern owing to the carcinogenicity of these compounds. Several organic compounds (amyrones, friedeline, aromatic A-noroleananes, syringaldehyde, vanillin, syringic acid, and vanillic acid) have been proposed as tracers of aerosols from biomass burning (Simoneit et al., 1996).

Other proposed tracers of biomass burning include the trace elements potassium and zinc (Andreae, 1983). These elements are enriched in biomass burning aerosols, and the ratios of K and Zn to black carbon (Cb) in aerosols from burning have been found to be quite constant (K/Cb ~1.3; Zn/Cb ~5.4‰, Cachier et al., 1996). Elemental analyses of coarse aerosols from prescribed savanna burns by Maenhaut et al. (1996) showed the fires were a major source for black carbon, P, K, Ca, Mn, Zn, Sr, and I. For fine particles ($r < 1$ μm), the flaming and smoldering phases of the fires were a major source for black carbon, Cl, Br, I, K, Cu, Zn, Rb, Sb, Cs, and Pb; and under flaming conditions also important for Na and S. These authors also found that fires could mobilize significant amounts of mineral dust, presumably through the convection associated with the fires.

Other Pollution-Derived Aerosols

Aerosols, both primary and secondary, are generated by a variety of pollution sources in addition to biomass burning; these include industrial processes, electric utilities, transportation, construction, and other fuel combustion. While large-scale urban air pollution is a consequence of modern industrial and technological development, smoke produced by indoor fires was perhaps the earliest form of air pollution (Brimblecombe, 1995). It is remarkable that the emissions from many anthropogenic sources are known with greater certainty than are those from natural sources. Even so, there are major gaps in our understanding of anthropogenic aerosol sources on a global scale (Graedel et al., 1993). These gaps include limited information on the geographical distribution of sources, inadequate measurements of the sizes of the particles emitted by the various sources, a lack of knowledge concerning the transformations of the particles as they age, and only a recent appreciation of the complex ways in which the entire mix of atmospheric constituents, especially nitrogen oxides, volatile organic compounds (VOCs), and ozone, affect the formation and composition of aerosols (Meng et al., 1997).

Pollution emissions are, of course, subject to many of the same processes discussed above for natural aerosol sources, such as new particle formation via gas-to-particle conversion or the condensation of volatile materials in plumes emitted by high-temperature sources. In addition studies of semivolatile organic compounds, including pollution-derived PAHs, have shown that the partitioning of these compounds between the gas phase and particles is largely determined by their subcooled liquid–vapor pressures (Bidleman and Foreman, 1987). Calculations by these authors based on the approach of Yamasaki et al. (1984) indicated that for

typical, urban, suspended-particle loads, 11 to 55% of the total mass of a substance with a vapor pressure of 10^{-6} Torr should be partitioned in the particulate phase. Other factors besides the total suspended particle load that can influence the vapor/particle partitioning of organic substance include relative humidity (Pankow et al., 1993), temperature, and the radiative flux (Kamens et al., 1988). Gas–particle partitioning studies of organics have shown that it is important to determine whether the gas-phase species are adsorbed to a solid particle's surface or absorbed into a liquid phase (Pankow, 1994). This same issue is certainly relevant to the gas–particle partitioning of inorganic species, specifically with respect to wetted aerosols and liquid droplets. Partitioning among the gas, liquid, and solid phases in the atmosphere is especially relevant for the chemistry of sulfur and nitrogen oxides and acid deposition, but further discussion of this topic is beyond the scope of this chapter.

Some quantitative estimates of trace element emissions from anthropogenic sources were produced in the 1970s and 1980s (e.g., Lantzy and Mackenzie, 1979; Nriagu, 1989; Pacyna, 1986; Nriagu and Pacyna, 1988). These emission estimates clearly show that the biogeochemical cycles of a number of trace elements have been severely perturbed by human activities (Table 2). However, of the trace elements it is atmospheric Pb that has been subject to the greatest perturbation, and while anthropogenic Pb has been spread throughout Earth, largely as a result of atmospheric transport (e.g., Murozumi et al., 1969; Patterson, 1987), the concentrations of Pb in the atmosphere have started to decline in response to the phase out of leaded gasolines (Huang et al., 1996).

There are examples of aerosol pollution even more extreme than Pb, and these involve the atmospheric releases of substances that exist purely as a result of human activities. Examples of the substances involved include synthetic organic chemicals, such as polychlorinated biphenyls (PCBs) and various types of pesticides. Radio-

TABLE 2 Percent of Total Atmospheric Emissions from Natural Sources[a]

Element	Percent from Natural Sources
Antimony	41
Arsenic	39
Cadmium	15
Chromium	59
Copper	44
Lead	4
Manganese	89
Mercury	41
Molybdenum	48
Nickel	35
Selenium	58
Vanadium	25
Zinc	34

[a]Data from Nriagu (1989).

active nuclides produced by nuclear weapons testing and by nuclear reactors also have been released into the environment, and these man-made nuclides also make up a component of the contemporary aerosol. The dispersal of these and other pollutant aerosols to the most remote parts of the globe is a measure of the efficiency with which atmospheric transport operates.

4 CONCLUDING REMARKS

Much of the current interest in aerosols focuses on two areas, first the connections between aerosols and climate, and second the links between aerosol pollution and human disease. Interest in the aerosol–climate connection centers on the direct and cloud-mediated effects of aerosols on solar radiation. Concern over the health effects of aerosol particles < 2.5 μm in diameter (the PM-2.5 fraction) is responsible for the recently enacted standards regarding particulate matter (PM) in the United States (Federal Register, 1997). Both of these general areas of interest require accurate information on the formation, composition, chemical reactivity, and transport of aerosol particles. Models and measurements have been used to determine where aerosols are produced and what they are made of, but advances in remote sensing and analytical methods will lead to a more comprehensive picture of the sources, composition, and reactivity of the atmospheric aerosol.

Beyond these concerns, it is now recognized that understanding the chemistry of aerosols is a key to dealing with other atmospheric constituents of immediate concern, including, for example, certain photochemical oxidants (Finlayson-Pitts and Pitts, 1997). Furthermore, understanding the linkages among the chemical cycles of aerosols, VOCs, and NO_x, will be required for the development of effective pollution control strategies (Meng et al., 1997). This newly recognized need for an integrated approach to understanding and controlling air pollution also will lead to improvements in the socioeconomic models that are becoming increasingly important in policy-making decisions.

Climate models are incorporating more and more information on the sources, composition, and fluxes of aerosols. Improved estimates of the quantities of gases and aerosols emitted into the atmosphere from natural and anthropogenic sources are being developed in association with the Global Emissions Inventory Activity (GEIA), a component of the International Global Atmospheric Chemistry Program (IGAC, e.g., Graedel et al., 1993). Given the global dimensions of aerosol pollution problems coupled with the heterogeneity of the aerosol distribution, it is appropriate that coordinated international efforts are being mounted to address both scientific and public health issues.

REFERENCES

An, S., T. S. Liu, Y. C. Lu, S. C. Porter, G. Kukla, W. H. Wu, and Y. M. Hua, The long-term paleomonsoon variation recorded by the loess-paleosol sequence in central China, *Quat. Int.*, *7/8*, 91–95, 1990.

Anderson, J. R., R. R. Buseck, T. L. Patterson, and R. Arimoto, Characterization of the Bermuda tropospheric aerosol by combined individual-particle and bulk-aerosol analysis, *Atmos. Environ.*, *30*, 319–338, 1996.

Andreae, M. O., Soot carbon and excess fine potassium: Long-range transport of combustion derived aerosols, *Science*, *220*, 1148–1151, 1983.

Arimoto, R., R. A. Duce, D. L. Savoie, J. M. Prospero, R. Talbot, J. D. Cullen, U. Tomza, N. F. Lewis, and B. J. Ray, Relationships among aerosol constituents from Asia and the North Pacific during PEM-West A, *J. Geophys. Res.*, *101*, 2011–2023, 1996.

Andreae, M. O., Climatic effects of changing atmospheric aerosol levels, in A. Henderson-Sellers (Ed.), *World Survey of Climatology*, Vol. 16: *Future Climates of the World*, Elsevier, Amsterdam, 1995, pp. 341–392.

Andreae, M. O., E. Atlas, H. Cachier, W. R. Cofer III, G. W. Harris, G. Helas, R. Koppmann, J-P. Lacaux, and D. E. Ward, Trace gas and aerosol emissions from savanna fires, in J. S. Levine (Ed.), *Biomass Burning and Global Change*, Vol. 1, MIT Press, Cambridge, MA, 1996, pp. 278–295.

Andreae, M. O., and P. J. Crutzen, Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry, *Science*, 276, 1052–1058, 1997.

Ballentine, D. C., S. A. Macko, V. C. Turekian, W. P. Gilhooly, and B. Martincigh, Chemical and isotopic characterization of aerosols collected during sugar cane burning in South Africa, in J. S. Levine (Ed), *Biomass burning and global change*, Vol. 1. MIT Press, Cambridge, MA, 1996, pp. 460–465.

Bandy, A. R., D. L. Scott, B. W. Blomquist, S. H. Chen, and D. C. Thornton, Low yields of SO_2 from dimethyl sulfide oxidation in the marine boundary layer, *Geophys. Res. Lett.*, *19*, 1125–1127, 1992.

Bates, T. S., J. A. Calhoun, and P. K. Quinn, Variations in the methanesulfonate to sulfate molar ratio in submicrometer marine aerosol particles over the South Pacific Ocean, *J. Geophys. Res.*, *97*, 9859–9865, 1992.

Baylor, E. R., V. Peters, and M. B. Baylor, Water-to-air transfer of virus, *Science*, *197*, 763–764, 1977.

Berresheim, H., Biogenic sulfur emissions from the Subantartic and Antartic oceans, *J. Geophys. Res.*, *92*, 13, 245–13, 1987.

Berresheim, H., P. H. Wine, and D. D. Davis, Sulfur in the atmosphere, in H. B. Singh (Ed.), *Composition, Chemistry and Climate of the Atmosphere*, Van Nostrand Reinhold, New York, 1995, pp. 251–307.

Betzer, P. R., K. L. Carder, R. A. Duce, J. T. Merrill, N. W. Tindale, M. Uematsu, D. K. Costello, R. W. Young, R. A. Breland, R. E. Bernstein, and A. M. Greco, Long-range transport of giant mineral aerosol particles, *Nature*, *336*, 568–571, 1988.

Bidleman, T. F., and W. T. Foreman, Vapor–particle partitioning of semivolatile organic compounds, in R. A. Hites and S. J. Eisenreich (Eds.), *Sources and Fates of Aquatic Pollutants*, American Chemical Society, Washington, DC, 1987, pp. 27–56.

Blanchard, D. C., The production, concentration, and vertical distribution of the sea–salt aerosol, *Ann. N.Y. Acad Sci.*, *338*, 330–347, 1980.

Blanchard, D. C., The production, distribution, and bacterial enrichment of the sea–salt aerosol, in P. S. Liss and W. G. N. Slinn (Eds.), *Air-Sea Exchange of Gases and Particles*, D. Reidel Publishing Company, Dordrecht, Holland, 1983, pp. 299–405.

Brimblecombe, P., History of air pollution, in H. B. Singh (Ed.), *Composition, Chemistry, and Climate of the Atmosphere*, Van Nostrand Reinhold, New York, 1995, pp. 1–18.

Brimblecombe, P., and S. L. Clegg, The solubility and behaviour of acid gases in the marine aerosol, *J. Atmos. Chem.*, 7, 1–18, 1988.

Buat-Ménard, P., H. Cachier, and R. Chesselet, Sources of particulate carbon in the marine atmosphere, in J. P. Riley, R. Chester, and R. A. Duce (Eds), *Chem. Oceanogr., Vol. 10*, Academic Press, London, 1989, 252–279.

Buat-Ménard, P., Global source strength and long–range atmospheric transport of trace elements emitted by volcanic activity, in A. H. Knap (Ed.), *The Long-Range Atmospheric Transport of Natural and Contaminant Substances*, Kluwer Academic, Dordrecht, 1990, pp. 163–175.

Cachier, H., C. Liousse, M.-H. Pertuisot, A. Gaudichet, F. Echaler, and J.-P. Lacaux, African fire particulate emission and atmospheric influence, in J. S. Levine (Ed), *Biomass burning and global change*, Vol. 1. MIT Press, Cambridge, MA, 1996, pp. 428–440.

Chameides, W. L., and A. W. Stelson, Aqueous–phase chemical processes in deliquescent sea-salt aerosols: A mechanism that couples the atmospheric cycles of S and sea salt, *J. Geophys. Res.*, *97*, 20565–20580, 1992.

Charlson, R. J., J. E. Lovelock, M. O. Andreae, and S. G. Warren, Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate, *Nature 326*, 655–661, 1987.

Clarke, A. D., J. L. Varner, F. Eisele, R. L. Mauldin, D. Tanner, and M. Litchy, Particle production in the remote marine atmosphere: Cloud outflow and subsidence during ACE–1, *J. Geophys Res.*, 103, 16397–16409, 1997.

Clegg, N. A., and R. Toumi, Sensitivity of sulphur dioxide oxidation in sea salt to nitric acid and ammonia gas phase concentrations, *J. Geophys. Res.*, 102, 23241–23249, 1997.

Covert, D. S., V. N. Kapustin, P. K. Quinn, and T. S. Bates, New particle formation in the marine boundary layer, *J. Geophys. Res.*, 97, 20581–20589, 1992.

Dentener, F. J., G. R. Carmichael, Y. Zhang, J. Lelieveld, and P. J. Crutzen, Role of mineral aerosol as a reactive surface in the global troposphere, *J. Geophys. Res., 101*, 22869–22889, 1996.

Dondero, Jr., T. J., R. C. Rendtorff, G. F. Mallison, R. M. Weeks, J. S. Levy, E. W. Wong, and W. Schaffner, An outbreak of Legionnaires' disease associated with a contaminated air-conditioning cooling tower, *N. Engl. J. Med, 302*. 365–370, 1980.

Duce, R. A., V. A. Mohnen, P. R. Zimmerman, D. Grosjean, W. Cautreels, R. Chatfield, R. Jaenicke, J. A. Ogren, E. D. Pellizzari, G. T. Wallace, Organic material in the global troposphere, *Rev. Geophys. Space Phys.*, *21*, 921–952, 1983.

Ericksson, E., The yearly circulation of chloride and sulphur in nature, meteorological, geochemical and pedological implications, Part 2, *Tellus*, *12*, 63–109, 1960.

Erickson, D. J. and R. A. Duce, On the global flux of atmospheric sea salt, *J. Geophys. Res.*, *93*, 14079–14088, 1988.

Federal Register, *62*, 38762–38896, 1997.

Finlayson–Pitts, B. J., M. J. Ezell, and J. N. Pitts, Formation of chemically active chlorine compounds by reactions of atmospheric NaCl particles with gaseous N_2NO_5 and $ClONO_2$, *Nature*, *337*, 241–244, 1989.

Finlayson–Pitts, B. J., and J. N. Pitts, Jr., Tropospheric air pollution: ozone, airborne toxics, polycyclic aromatic hydrocarbons, and particles, *Science*, *276*, 1045–1052, 1997.

Gillette, D. A., I. H. Blifford, and D. W. Fryrear, The influence of wind velocity on the size distributions of aerosols generated by the wind erosion of soils, *J. Geophys. Res.*, *79*, 4068–4075, 1974.

Glaccum, R. A., and J. M. Prospero, Saharan aerosols over the tropical North Atlantic-mineralogy, *Mar. Geol.*, *37*, 295–321, 1980.

Graedel, R. E., T. S. Bates, A. F. Bouwman, D. Cunnold, J. Dignon, I. Fung, D. J. Jacob, B. K. Lamb, J. A. Logan, G. Marland, P. Middleton, J. M. Pacyna, M. Placet, and C. Veldt, A compilation of inventories of emissions to the atmosphere, *Global Biogeochem. Cycles*, *7*, 1–26, 1993.

Hao, W. M., and M. H. Liu, Spatial and temporal distribution of tropical biomass burning, *Global Biogeochem. Cycles*, *8*, 495–503, 1994.

Hegg, D. A., L. F. Radke, and P. V. Hobbs, Particle production associated with marine clouds, *J. Geophys. Res.*, *95*, 13917–13926, 1990.

Heintzenberg, J., Fine particles in the global troposphere - a review, *Tellus*, *41B*, 149–160, 1989.

Hickey, J. L. S., and P. C. Reist, Health significance of airborne microorganisms from wastewater treatment processes, *J. Water Pollut. Control Fed.*, *47*, 2758–2773, 1975.

Hoogheimstra, H., Variations of the NW African trade wind regime during the last 140,000 years: Changes in pollen flux evidenced by marine sediment records, in M. Leinen and M. Sarnthein (Eds.), *Paleoclimatology and Paleometeorology: Modern and Past Patterns of Global Atmospheric Transport*, NATO ASI Series, Kluwer Academic, Dordrecht, 1987, pp. 733–770.

Huang, S., R. Arimoto, and K. Rahn, Changes in atmospheric lead and other pollution-derived trace elements at Bermuda, *J. Geophys. Res.*, *101*, 21033–21040, 1996.

Husar, R. B., J. M. Prospero and L. L. Stowe. Characterization of tropospheric aerosols over the oceans with the NOAA/AVHRR optical thickness operational product. *J. Geophys. Res. 102*, 16, 16889–16909, 1997.

Imshenetsky, A. A., S. V. Lysenko, G. A. Kazakov, Upper boundary of the biosphere, *Appl. Environ. Microbiol.*, *35*, 1–5, 1978.

Johnson, L. R., Particle-size fractionation of eolian dusts during transport and sampling, *Marine Geo.*, *21*, M17–M21, 1976.

Junge, C. E., *Air Chemistry and Radioactivity*, Academic Press, New York, 1963.

Kamens, R. M., Z. Guo, J. N. Fulcher, and D. A. Bell, Influence of humidity, sunlight, and temperature on the daytime decay of polyaromatic hydrocarbons on atmospheric soot particles, *Environ. Sci. Technol.*, *22*, 103–108, 1988.

Kawamura, K., and K. Usukura, Distributions of low molecular weight dicarboxylic acids in the North Pacific aerosol samples, *J. Oceanogr.*, *49*, 271–283, 1993.

Keene, W. C., A. A. P. Pszenny, D. J. Jacob, R. A. Duce, J. N. Galloway, J. J. Schultz–Tokos, H. Sievering, and J. Boatman, The geochemical cycling of reactive chlorine through the marine troposphere, *Global Biogeochem. Cycles*, *4*, 407–430, 1990.

Keene, W. C., R. Sander, A. A. P. Pszenny, R. Vogt, P. J. Crutzen, and J. N. Galloway, Aerosol pH in the marine boundary layer: A review and model evaluation, *J. Aerosol Sci.*, 29, 239–356, 1998.

Koga, S. H. Tanaka, M. Yamato, T. Yamanouchi, F. Nishio and Y. Iwasaka, Methanesulfonic acid and non-sea-salt sulfate over both hemispheric oceans, *J. Meteorol. Soc. Jpn.*, *69*, 1–14, 1991.

Lambert, G., M.-F. Le Cloarec, and M. Pennisi, Volcanic output of SO_2 and trace metals: A new approach, *Geochim. Cosmochim. Acta*, 52, 39–42, 1988.

Lambert, G., G. Polian, J. Sanak, A. Buisson, R. Ardouin, and A. Jegou, Volcanic output of long-lived radon daughters, *J. Geophys. Res.*, *87*, 11103–11108, 1982.

Lantzy, R. L., and F. T. Mackenzie, Global cycles and assessment of man's impact, *Geochim. Cosmochim. Acta*, 43, 511–515, 1979.

Leathers, C. R., Plant components of desert dust in Arizona and their significance for man, in *Geological Society of America Special Paper 186*, Desert Dust: Origin, Characteristics, and Effect on Man, T. L. Péwé (Ed.), 1981, pp. 191–206.

Lee, D. S., J. M. Edmond, and K. W. Bruland, Bismuth in the Atlantic and North Pacific: A natural analogue to plutonium and lead? *Earth Planet. Sci. Lett.*, *76*, 254–2262, 1986.

Levetin, E., Fungi, in H. A. Burge (Ed.), *Bioaerosols*, Lewis Publishers, Boca Raton, 1995, pp. 87–120.

Lin, X., and W. L. Chameides, CCN formation from DMS oxidation without SO_2 acting as an intermediate, *Geophys. Res. Lett.*, *20*, 579–582, 1993.

Liu, T., et al (Thirty-five coauthors), *Loess and the Environment*, China Ocean Press, Beijing, 251 pp, 1985.

MacIntyre, F., and J. W. Winchester, Phosphate ion enrichment in drops from breaking bubbles, *J. Phys. Chem.*, *73*, 2163–2169, 1969.

MacIntyre, F., Chemical fractionation and sea-surface microlayer processes, in E. D. Goldberg (Ed.), *The Sea*, Vol. 5, J. Wiley, New York, 1974, pp. 245–299.

Maenhaut, W., I. Salma, J. Cafmeyer, H. J. Annegarn, and M. O. Andreae, Regional atmospheric aerosol composition and sources in the eastern Transvaal, South Africa, and impact of biomass burning, *J. Geophys. Res.*, *101*, 23631–23650, 1996.

Mamane, Y., and J. Gottlieb, Nitrate formation on sea-salt and mineral particles—a single particle approach, *Atmos. Environ.*, *26A*, 1763–1769, 1992.

Mazurek, M., C. Laterza, L. Newman, P. Daum, W. R. Cofer III, J. S. Levine, and E. L. Winstead, Composition of carbonaceous smoke particles from prescribed burning of a Canadian boreal forest: Organic aerosol characterization by gas chromatography, in S. Levine (Ed.), *Biomass Burning and Global Change*, Vol. 2, MIT Press, Cambridge, MA, 1996, pp. 840–847.

McCormick, M. P., L. W. Thomason, and C. R. Trepte, Atmospheric effects of the Mt. Pinatubo eruption, *Nature*, *373*, 399–404, 1995.

Meng, Z., D. Dabdub, and J. H. Seinfeld, Chemical coupling between atmospheric ozone and particulate matter, *Science*, *277*, 116–119, 1997.

Mouri, H., K. Okada, and K. Shigehara, Variation of Mg, S, K and Ca Contents in individual sea-salt particles, *Tellus*, *45B*, 80–85, 1993.

Muilenberg, M. G., The outdoor aerosol, in H. A. Burge, (Ed.), *Bioaerosols*, Lewis Publishers, Boca Raton, 1995, pp. 163–204.

Murozumi, M., T. J. Chow, and C. Patterson, Chemical concentrations of pollutant lead aerosols, terrestrial dusts and sea salts in Greenland and Antarctic snow strata, *Geochim. Cosmochim. Acta*, *33*, 1247–1294, 1969.

Nriagu, J. O., A global assessment of natural sources of atmospheric trace metals, *Nature*, *338*, 47–49, 1989.

Nriagu, J. O., and J. M. Pacyna, Quantitative assessment of worldwide contamination of air, water and soils by trace metals, *Nature*, *333*, 134–139, 1988.

Okada, K., Y. Ishizaka, T. Masuzawa, and K. Isono, Chlorine deficiency in coastal aerosols, *J. Meteorol. Soc. Jpn.*, *56*, 501–507, 1978.

Pacyna, J. M., Atmospheric trace elements from natural and anthropogenic sources, in J. O. Nriagu and C. I. Davidson (Eds.), *Toxic Metals in the Atmosphere*, Wiley, New York, 1986, pp. 33–52.

Pankow, J. F., An absorption model of gas/particle partitioning of organic compounds in the atmosphere, *Atmos. Environ.*, *28*, 185–188, 1994.

Pankow, J. F., J. M. E. Storey, and H. Yamasaki, Effects of relative humidity on gas/particle partitioning of semivolatile organic compounds to urban particulate matter, *Environ. Sci. Technol.*, *27*, 2220–2226, 1993.

Parungo, F., Z. Li, X. Li, D. Yang, and J. Harris, Gobi dust storms and the Great Green Wall, *Geophys. Res. Lett.*, *21*, 999–1002, 1994.

Parungo, F. P., C. T. Nagamoto, and J. M. Harris, Temporal and spatial variations of marine aerosols over the Atlantic Ocean, *Atmos. Res.*, *20*, 23–37, 1986.

Patterson, C., Global pollution measured by lead in mid-ocean sediments, *Nature*, *326*, 244–245, 1987.

Peltzer, E. T., and R. B. Gagosian, Organic geochemistry of aerosols over the Pacific Ocean, in J. P. Riley, R. Chester, and R. A. Duce, (Eds.), *Chemical Oceanography*, Academic, London, 1989, pp. 281–338.

Perry, K. D., and P. V. Hobbs, Further evidence for particle nucleation in clean air adjacent to marine cumulus clouds, *J. Geophys. Res.*, *99*, 22803–22818, 1994.

Péwe, T. L., Desert dust: An overview, in *Geological Society of America Special Paper 186*, 1981, pp. 1–10.

Pimentel, D., C. Harvey, P. Resosudarmo, K. Sinclair, D. Kurz, M. McNair, S. Crist, L. Shpritz, L. Fitton, R. Saffouri, and R. Blair, Environmental and economic costs of soil erosion and conservation benefits, *Science 267*, 1117–1123, 1995.

Polian, G., and G. Lambert, Radon daughters and sulfur output from Erebus Volcano, Antarctica, *J. Volcanol. Geotherm. Res.*, *6*, 125–137, 1979.

Pósfai, M., J. R. Anderson, and P. R. Buseck, Compositional variations of sea-salt-mode aerosol particles from the North Atlantic, *J. Geophys. Res.*, *100*, 23063–23074, 1995.

Prospero, J. M. and R. T. Nees, Dust concentration of the equatorial North Atlantic: possible relationship to the Sahelian drought, *Science*, *196*, 1196–1198, 1977.

Pszenny, A. A. P., A. J. Castell, J. N. Galloway, and R. A. Duce, A study of the sulfur cycle in the Antartic marine boundary layer, *J. Geophys. Res.*, *94*, 9819–9830, 1989.

Pye, K., *Aeolian dust and dust deposits*, Academic Press, London, 1987, 334 pp.

Rahn, K., The chemical composition of the atmospheric aerosol, Tech. Report, Grad. Sch. of Oceanogr., Univ. of Rhode Island, Kingston, 265 pp. 1976.

Savoie, D. L., and J. M. Prospero, Comparison of oceanic and continental sources of non-seasalt sulphate over the Pacific Ocean, *Nature*, *339*, 685–687, 1989.

Savoie, D. L., J. M. Propero, R. Arimoto, and R. A. Duce, Nonsea-salt sulfate and methanesulfonate at American Samoa, *J. Geophys. Res.*, *99*, 3587–3596, 1994.

Schlesinger, W. H., J. F. Reynolds, G. L. Cunningham, L. F. Huenneke, W. M. Jarrell, R. A. Virginia, and W. G. Whitford, Biological feedbacks in global desertification, *Science*, *247*, 1043–1048, 1990.

Schroeder, W. H., and P. Urone, Formation of nitrosyl chloride from sea particles in air, *Environ. Sci. Technol.*, *8*, 756–758, 1974.

Schütz, L., and K. A. Rahn, Trace-element concentrations in erodible soils, *Atmos. Environ.*, *16*, 171–176, 1982.

Sholkovitz, E. R., T. M. Church, and R. Arimoto, Rare earth element composition of rainwater, wet deposition and aerosols, *J. Geophys. Res.*, 98, 20587–20599, 1993.

Sievering, H., J. Boatman, E. Gorman, Y. Kim, L. Anderson, G. Ennis, M. Luria, and S. Pandis, Removal of sulphur from the marine boundary layer by ozone oxidation in sea-salt aerosols, *Nature*, *360*, 571–573, 1992.

Sievering, H., E. Gorman, T. Ley, A. Pszenny, M. Springer-Young, J. Boatman, Y. Kim, C. Nagamoto, and D. Wellman, Ozone oxidation of sulfur in sea-salt aerosol particles during the Azores Marine Aerosol and Gas Exchange experiment, *J. Geophys. Res.*, *100*, 23075–23081, 1995.

Simoneit, B. R. T., M. R. bin Abas, G. R. Cass, W. F. Rogge, M. A. Mazurek, L. J. Standley, and L. M. Hildermann, Natural organic compounds as tracers for biomass combustion in aerosols, in S. Levine (Ed.), *Biomass Burning and Global Change*, Vol. 1, MIT Press, Cambridge, MA, 1996, pp. 509–517.

Swap, R., M. Garstang, S. Greco, R. Talbot, and P. Kållberg, Saharan dust in the Amazon Basin, *Tellus*, *44B*, 133–149, 1992.

Symonds, R. B., W. I. Rose, and M. H. Reed, Contribution of Cl- and F-bearing gases to the atmosphere by volcanoes, *Nature*, *334*, 415–418, 1988.

Taylor, S. R. and S. M. McLennan, *The Continental Crust: Its Composition and Evolution*, 312 pp., Blackwells, Oxford, England, 1985.

Tegen, I., A. A. Lacis, and I. Fung, The influence on climate forcing of mineral aerosols from disturbed soils, *Nature*, *380*, 419–423, 1996.

Wallace, G. T., and R. A. Duce, Concentration of particulate trace metals and particulate organic carbon in marine surface waters by a bubble flotation mechanism, *Marine Chem.*, *2*, 157–181, 1975.

Weisel, C. P., R. A. Duce, J. L. Fashing, and R. W. Heaton, Estimates of the transport of trace elements from the ocean to the atmosphere, *J. Geophys. Res.*, *89*, 11607–11618, 1984.

Whitby, K. T., The physical characteristics of sulfur aerosols, *Atmos. Environ.*, *12*, 135–159, 1978.

Winchester, J. W., and M.-X. Wang, Acidic sulfur uptake by alkaline dust in the Asia-Pacific region, in L. Newman, W. Wang, and C. S. Kiang (Eds.), *Proceedings International Conference on Global and Regional Environmental Atmospheric Chemistry*, U.S. Department of Energy, Washington, DC, 1990, pp. 13–23.

Wu, P.-M., and K. Okada, Nature of coarse nitrate particles in the atmosphere—a single particle approach, *Atmos. Environ.*, *28*, 2053–2060, 1994.

Yamasaki, H., K. Kuwata, and Y. Kuge, Determination of vapor pressure of polycyclic aromatic hydrocarbons in the supercooled liquid phase and their adsorption on airborne particulate matter, *Nippon Kagaku Kaishi*, *8*, 1324–1329 (*Chem. Abstr.*, *101*, 156747p), 1984.

Zhou, G., and K. Tazaki, Seasonal variation of gypsum in aerosol and its effect on the acidity of wet precipitation on the Japan Sea side of Japan, *Atmos. Environ.*, *30*, 3301–3308, 1996.

Zoller, W. H., J. R. Parrington, and J. M. P. Kotra, Iridium enrichment in airborne particles from Kilauea Volcano: January 1983, *Science*, *222*, 1118–1121, 1983.

CHAPTER 12

AEROSOLS: FORMATION AND MICROPHYSICS IN THE TROPOSPHERE

JOHN H. SEINFELD

1 INTRODUCTION

Particles in the atmosphere arise from natural sources, such as wind-borne dust, sea spray, and volcanoes, and from anthropogenic activities, such as combustion of fuels. While an aerosol is technically defined as a suspension of fine solid or liquid particles in a gas, common usage refers to the aerosol as the particulate component only. Emitted directly as particles (primary aerosol) or formed in the atmosphere by gas-to-particle conversion processes (secondary aerosol), atmospheric aerosols are generally considered to be the particles that range in size from a few nanometers to tens of micrometers in diameter. Once airborne, particles can change their size and composition by condensation of vapor species or by evaporation, by coagulating with other particles, by chemical reaction, or by activation in the presence of water supersaturation to become fog and cloud droplets. Particles smaller than 1 μm diameter generally have atmospheric concentrations in the range from around tens to thousands per cubic centimeter; those exceeding 1 μm diameter are usually found at concentrations less than 1 per cm^3.

A significant fraction of the tropospheric aerosol is anthropogenic in origin. Chemical components of tropospheric aerosols include sulfate, ammonium, nitrate, sodium, chloride, trace metals, carbonaceous material, crustal elements, and water. The carbonaceous fraction consists of both elemental and organic carbon. Elemental carbon, also called black carbon, graphitic carbon, or soot, is emitted directly into the atmosphere, predominantly from combustion processes. Particulate organic

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

carbon is emitted directly by sources or can result from atmospheric condensation of low-volatility organic gases.

2 PARTICLE SIZE DISTRIBUTION

Size is the most important single characteristic of an aerosol particle. For a spherical particle, diameter (or radius) is the usual reported dimension. When a particle is not spherical, the size can be reported either in terms of a length scale characteristic of its silhouette or of a hypothetical sphere with equivalent dynamic properties, such as settling velocity in air. For example, the *aerodynamic diameter* of a particle represents the diameter of a unit density ($\rho_p = 1$ g/cm^3) sphere having the same terminal settling velocity as the particle sampled, whatever its size, shape, or density.

When particles, at total number concentration N (particles/cm^3), are measured and the number of particles dN having diameters between D_p and $D_p + dD_p$, where dD_p is a small increment of diameter, are counted, the particle size distribution $n(D_p)$ is defined as $n(D_p) = dN/dD_p$ (reciprocal micrometers per cubic centimeters), where D_p is usually measured in micrometers. The integral of the size distribution over all sizes is the total number concentration:

$$N = \int_0^\infty n(D_p)\, dD_p \tag{1}$$

The log-normal distribution is particularly useful for representing aerosol size distributions because it does not allow negative particle sizes,

$$n(D_p) = \frac{N}{\sqrt{2\pi}\ln\sigma_g}\exp\left[-\frac{(\ln D_p - \ln D_g)^2}{2\ln^2\sigma_g}\right] \tag{2}$$

where D_g is the geometric mean diameter and σ_g is the geometric standard deviation. These parameters can be determined from discrete particle count data by

$$\ln D_g = \frac{1}{N}\sum_i N_i \ln D_{pi} \tag{3}$$

$$\ln\sigma_g = \left[\frac{1}{N}\sum_i (\ln D_{pi} - \ln D_g)^2\right]^{1/2} \tag{4}$$

3 RESIDENCE TIMES OF PARTICLES IN THE TROPOSPHERE

Particles are eventually removed from the atmosphere by two mechanisms: deposition at Earth's surface, so-called dry deposition, and scavenging by droplets, so-called wet deposition (Seinfeld and Pandis, 1998). Because wet and dry deposition lead to relatively short residence times in the troposphere and because the geographical distribution of particle sources is highly nonuniform, tropospheric aerosols

vary widely in concentration and composition over Earth. Whereas atmospheric trace gases have lifetimes ranging from less than a second to a century or more, the residence times of particles in the troposphere vary only from a few days to a few weeks.

The dry deposition flux of particles to the surface, F_d, is assumed to be proportional to the particle concentration at a reference height, C, i.e., $F_d = v_d C$, where the proportionality constant v_d, the deposition velocity, depends on the meteorological state of the atmosphere and the size of the particles. Three processes serve to deliver particles to Earth's surface: gravitational settling, turbulent transport, and Brownian diffusion. Although virtually any atmospheric flow is turbulent, a very thin laminar sublayer exists immediately adjacent to the surface. Turbulence brings particles down to the laminar sublayer, through which Brownian diffusion and settling govern transport. Small particles have a relatively large Brownian diffusivity, so move efficiently through the sublayer, whereas larger particles transfer primarily via inertia or settling. Those in between, in the size range of 0.1 to 1 μm diameter, are deposited about an order of magnitude slower than those at either the small or large extremes because none of the mechanisms is relatively effective in this intermediate size range.

Wet deposition involves the scavenging of particles by droplets and the subsequent removal by precipitation. Scavenging is necessary, but not sufficient, for wet deposition to occur since cloud or rain drops can evaporate, and if this occurs, the scavenged particle is returned to the air mass. As opposed to dry deposition, which operates only at Earth's surface, wet deposition serves to remove particles from the entire air mass. The rate of particle collection by falling drops is proportional to the number of drops, their settling velocity, their cross-sectional area, and a collection efficiency. The efficiency with which a particle is collected by a falling drop depends on the mechanics of particle motion in the vicinity of the drop. As with dry deposition, there is a minimum in the total collection efficiency in the 0.1 to 1 μm size range—small particles diffuse to the drop surface, larger ones collide with it, while in between neither process is very efficient.

Particles that become activated to grow to fog or cloud droplets are termed cloud condensation nuclei (CCN). At a given mass of water-soluble material in the particle, there is a critical value of the ambient water supersaturation, above which the particle undergoes an unstable process of spontaneous water accretion, leading to a cloud droplet (Seinfeld and Pandis, 1998). The critical water supersaturation for activation results from a combination of the curvature increase in and the solute concentration lowering of the water vapor pressure over a droplet. The number of particles that can act as CCN thus depends on the water supersaturation. For marine stratiform clouds, for example, supersaturations are in the range of 0.1 to 0.5%, which corresponds to a minimum CCN particle diameter of 0.05 to 0.14 μm. CCN number concentrations vary from fewer than $100/\text{cm}^3$ in remote marine regions to a few thousand per cubic centimeter in polluted urban areas. Once activated, fog and cloud droplets grow to sizes exceeding 10 μm diameter. Particles that are not activated to form droplets may remain as airborne aerosol or be removed by falling drops.

Aerosol lifetimes in the atmosphere depend primarily on the size of the particle and the height in the atmosphere at which the particle resides. The residence time τ can be viewed as an exponential half-life, the time required for a population of particles of a given size to decay to $1/e$ of its initial concentration.

An empirical expression for atmospheric particle residence time as a function of particle size and altitude that is useful for estimates is (Jaenicke, 1988)

$$\frac{1}{\tau} = \frac{1}{K}\left[\left(\frac{D_p}{D_{\max}}\right)^2 + \left(\frac{D_p}{D_{\max}}\right)^{-2}\right] + \frac{1}{\tau_{\text{wet}}} \tag{5}$$

where $K = 1.28 \times 10^8$ s (constant), $D_{\max} = 0.6$ μm (the diameter of particle with maximum residence time), and τ_{wet} is the lifetime for removal of particles by wet deposition. The first term on the right-hand side of Eq. (5) represents dry removal at Earth's surface; τ_{wet} depends mainly on the altitude in the atmosphere. Roughly three altitude regions can be distinguished by the frequency with which precipitation scavenging occurs:

Height $\leq$ 1.5 km (lower troposphere)	$\tau_{\text{wet}} \approx 6.9 \times 10^5$ s (8 days)
Middle troposphere to tropopause	$\tau_{\text{wet}} \approx 1.8 \times 10^6$ s (3 weeks)
Tropopause and above	$\tau_{\text{wet}} \approx 1.7 \times 10^7$ s (200 days)

Figure 1 shows atmospheric particle residence time τ as a function of particle radius, $D_p/2$. Dry removal predominates for particle radii either much smaller or larger than 0.3 μm; in the region around 0.3 μm wet scavenging is the most effective removal mechanism.

4 TROPOSPHERIC AEROSOLS

The relatively short residence time of aerosols in the troposphere, usually less than a couple of weeks, results in significant spatial variations in particle concentration, size, and composition. In an effort to categorize tropospheric aerosols, eight approximate classes can be identified: marine, remote continental, nonurban continental, urban, desert, polar, biomass burning, and background (free troposphere) (Heintzenberg, 1989; Fitzgerald, 1991; Jaenicke, 1993).

Marine Aerosol

Particles over the remote oceans are largely of marine origin (Savoie and Prospero, 1989). Marine atmosphere particle concentrations are normally in the range of 100 to 300/cm^3 (Fitzgerald, 1991). Typically the coarse particle mode (diameters exceeding 1 μm), comprising 95% of the total mass but only 5 to 10% of the particle number, results from the evaporation of sea spray produced by bursting bubbles or wind-induced wave breaking (Blanchard and Woodcock, 1957; Monahan et al.,

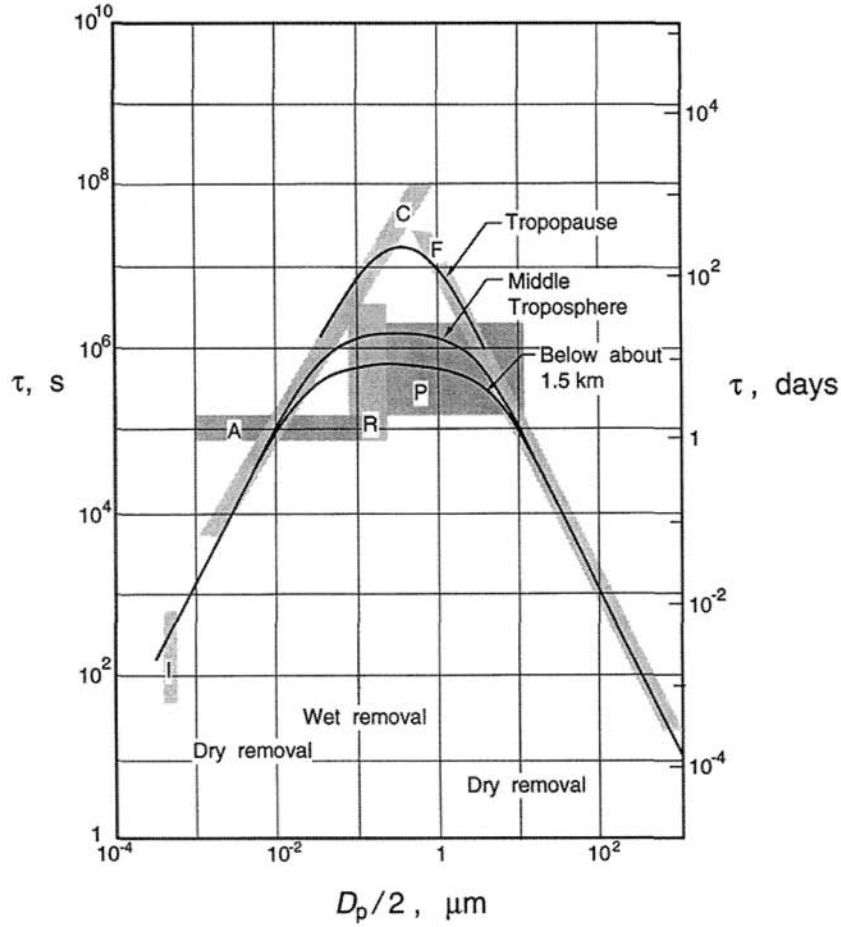


Figure 1 Residence times of tropospheric aerosols as a function of particle radius $D_p/2$. I, small atmospheric ions; A, so-called Aitken particles, radii 0.001 to 0.1 μm, residence time estimated from geographical distributions; C, residence time calculated based on Brownian coagulation of particles; R, radioactivity; P, precipitation removal; F, sedimentation. The three curves shown correspond to the three altitude levels indicated based on Eq. (5). Adapted from Jaenicke (1988).

1983). Typical sea salt aerosol concentrations in the marine boundary layer (MBL) are thought to be around 20 to 30/cm^3 (Blanchard and Cipriano, 1987; O'Dowd and Smith, 1993).

Remote Continental Aerosol

Aerosol number concentrations in remote continental regions average around 10^4/cm^3, and mass concentrations average about 20 μg/m^3 (Bashurova et al.,

1992; Koutsenogii et al., 1993; Koutsenogii and Jaenicke, 1994). A typical remote continental aerosol number distribution has three modes centered at diameters about 0.02, 0.12, and 1.8 μm (Bashurova et al., 1992; Koutsenogii et al., 1993; Koutsenogii and Jaenicke, 1994).

Urban Aerosol

Urban aerosols are strongly anthropogenic in origin, with a definite combustion signature. Number concentrations usually exceed $10^5/cm^3$, and size distributions typically exhibit three modes: named nuclei, accumulation, and coarse. The constituents of urban aerosol comprise the full spectrum of compounds possible in the atmospheric aerosol: sulfate, nitrate, ammonium, elemental and organic carbon (EC and OC), and crustal compounds (silicon, aluminum, calcium, and iron oxides). These aerosols result from primary emissions (EC, OC, soil material) and gas-to-particle transformation of the oxides of nitrogen, hydrocarbons, ammonia, and SO_2.

Desert Aerosol

Large amounts of dust are emitted to the atmosphere from deserts, especially during high wind periods. Most of these particles are coarse ($D_p > 1$ μm), and many are deposited close to their source; some fraction of the smaller particles can be transported over large distances (Prospero, 1990). For example, dust from the Sahara is regularly detected on Barbados Island across the Atlantic Ocean (Andreae, 1995). The chemical composition of desert aerosol reflects its soil source and is often rich in calcium compounds and other alkaline elements.

Polar Aerosol

Air masses generally remain over polar ice for extended periods of time. Any large particles that are present have sufficient time to deposit out, leaving a monodisperse aerosol with a mean size of about 0.15 μm and a number concentration in the range of 15 to 150 particles/cm^3 (Browell et al., 1992). Suitable meteorological conditions leading to the transport of anthropogenic aerosol to high latitudes in winter and early spring produce so-called arctic haze (Barrie, 1986).

Background Aerosol

The aerosol in the mid and upper troposphere, the so-called free troposphere, is often termed background aerosol. It is well-aged aerosol with a composition and size distribution reflecting the simultaneous effects of gas-to-particle conversion, long-range transport, and removal processes. The number concentration of background aerosol is in the range of 300/cm^3 (Raes et al., 2000), and its size distribution is nearly monodisperse with peak diameters in the 0.2 to 0.5 μm range (Leaitch and Isaac, 1991). Regions with the lowest mass concentrations generally exhibit the highest number concentrations (Clarke, 1993), suggesting that nucleation may be

a major source of particle number. Volatility measurements of the free tropospheric aerosol suggest a composition dominated by sulfates (Clarke, 1993; Hofmann, 1993)

Biomass Burning Aerosol

Remote biomass burning (forest and savanna fires, agricultural burning, etc.) is a major source of both primary (ash, elemental carbon) and secondary (organic carbon, sulfate, nitrate, and ammonium) aerosol. The chemical composition of the aerosol produced depends on the characteristics of the combustion: hot, flaming fires (e.g., savanna fires) emit mainly EC aerosol, while smoldering fires emit mainly organic particles (Andreae, 1995). The number concentrations of aerosols produced by biomass fires are of the order of tens of thousands of particles per cubic centimeter close to the source and less than $1000/cm^3$ after a few days of transport. Mass median diameters in fresh fire plumes are typically in the range of 0.1 to 0.3 μm and evolve toward values in the range of 0.2 to 0.4 μm during the first few hours after emission.

Nonurban Continental Aerosol

Nonurban continental aerosol is often acidic as a result of anthropogenic sulfate or nitrate. Typical aerosol number concentrations of nonurban continental aerosol are in the range of $10^3/cm^3$, with mass concentrations around 30 $\mu g/m^3$ (Anderson et al., 1993). The aerosol mass distribution usually exhibits a trimodal structure similar to that of urban aerosol.

Figure 2 shows approximate atmospheric aerosol number concentration N and volume concentration V as a function of altitude z. Because of the large variations in number and volume concentrations in the lower troposphere (remote continental, rural continental, urban, marine, polar), the vertical profiles fan out at low altitudes

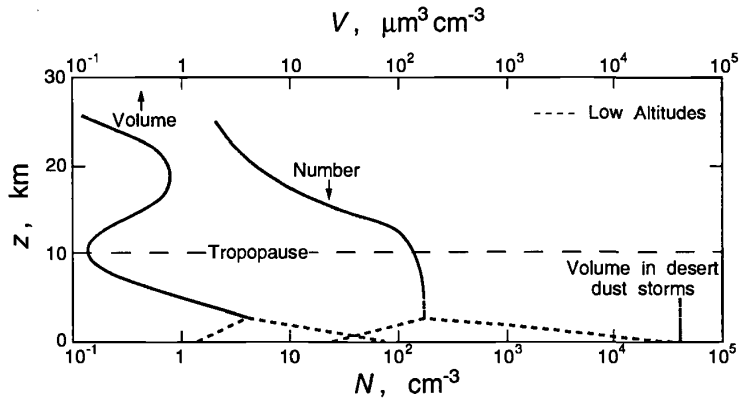


Figure 2 Atmospheric aerosol number N and volume V concentrations as a function of altitude z. Adapted from Jaenicke (1988).

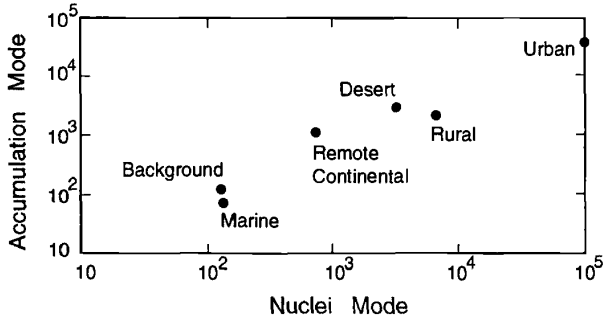


Figure 3 Typical aerosol number concentrations in accumulation and nuclei modes for six classes of global aerosols. Clement and Ford (1997), based on number concentrations reported by Jaenicke (1993). Reprinted from *Journal of Aerosol Science*, Vol. 28, No. 1, C. F. Clement and I. J. Ford, Properties and modelling of global aerosols, 5743-4, Copyright 1997, with permission from Elsevier Science.

(dashed lines). Aerosol number concentration decreases continuously with increasing altitude, whereas volume concentration decreases up to the tropopause (about 10 km) and then increases in the stratospheric aerosol layer, reaching maximum between 15 and 20 km altitude.

Atmospheric aerosol number concentrations can be naturally divided into three groups: nuclei mode particles with diameters ≤ 0.01 μm, accumulation mode particles with diameters 0.01 to 1 μm, and coarse particles with diameters exceeding 1 μm. Figure 3 shows the total number concentrations of the nuclei and accumulation modes for urban, rural, desert, remote continental, marine, and background aerosols. (There is no obvious nucleation mode for the polar aerosol.) Number concentrations of the two modes are seen to be equal over a range of 3 orders of magnitude in different parts of the atmosphere.

5 AEROSOL MICROPHYSICS

The atmosphere subjects aerosol particles to an array of transport and transformation processes that alter their size, number, and composition. Advective and turbulent transport of particles is nearly identical to that of the interstitial gas. Transformation processes include condensation and evaporation, which result from diffusion of vapors between the particle and the interstitial gas, homogeneous nucleation to produce new particles from supersaturated vapors, coagulation, which combines two particles into one by collision and sticking, and chemical reactions occurring in individual particles (Seinfeld and Pandis, 1998). That a major portion of atmospheric aerosol mass is secondary in nature is indicative of the importance of gas-to-particle conversion.

The aqueous phase of atmospheric aerosols contains primarily strong electrolytes such as sodium chloride, nitric and sulfuric acids, and ammonium. At relative

humidities much below saturation, the vast majority of water in the atmosphere is in the vapor phase, and therefore any liquid water associated with aerosol particles is too small to affect the ambient relative humidity. For relative humidities below saturation, water is in equilibrium between the vapor and aqueous phases because the characteristic time for water equilibration is relatively short compared to all other processes taking place. Other volatile aerosol species may or may not be in equilibrium depending on their equilibration characteristic time (Seinfeld and Pandis, 1998).

Much nucleation research relevant to the atmosphere has been focused, via measurement and theory, on the binary nucleation of sulfuric acid and water (Seinfeld and Pandis, 1998). Although a classical theory of binary nucleation of sulfuric acid and water exists (Jaecker-Voirol and Mirabel, 1989), substantial uncertainty still remains as to how accurately this classical theory represents the actual nucleation process. Measured nucleation rates can differ from theoretically predicted values by several orders of magnitude. From the point of view of atmospheric applications, significant nucleation rates can be defined as those exceeding 1 nucleus/cm^3 s.

Coagulation is the process whereby two particles collide and stick to form a single particle. Atmospheric processes that may lead to particle collisions include Brownian motion, turbulent shear, and differential settling. The latter two can be shown to be much less effective in this regard than Brownian motion (Wexler et al., 1994). In atmospheric aerosol dynamics, coagulation does not play a significant role unless number concentrations are relatively high and/or residence times are relatively long.

6 CONCLUSION

Despite significant progress in our understanding of the global aerosol system over the past two decades, our knowledge of the sources and dynamics of atmospheric aerosols remains limited. Difficulties associated with measurement of the aerosol size/composition distribution, combined with the significant spatial and temporal variability of tropospheric aerosol, have resulted in only scattered knowledge of its global distribution. Aerosols in remote locations and in the middle and upper troposphere have received relatively little attention, and their size/composition distribution remains largely unexplored. Most of the existing measurements are ground based, and, consequently, there is a general lack of information on the vertical aerosol distribution. Moreover, few measurements of the chemical composition of the smallest atmospheric particles, e.g., smaller than 50 nm diameter, are available.

REFERENCES

Anderson, B. E., G. L. Gregory, J. D. W. Barrick, J. E. Collins, G. W. Sachse, D. Bagwell, M. C. Shipham, J. D. Bradshaw, and S. T. Sandholm, The impact of U.S. continental outflow on

ozone and aerosol distributions over the western Atlantic, *J. Geophys. Res.*, *98*, 23477–23489, 1993.

Andreae, M. O., Climate effects of changing atmospheric aerosol levels, in A. Henderson-Sellers (Ed.), *World Survey of Climatology*, Elsevier, Amsterdam, 1995, pp. 347–398.

Barrie, L. A., Arctic air pollution: An overview of current knowledge, *Atmos. Environ.*, *29*, 643–663, 1986.

Bashurova, V. S., V. Dreiling, T. V. Hodger, R. Jaenicke, K. P. Koutsenogii, P. K. Koutsenogii, M. Kraemer, V. I. Makarov, V. A. Obolkin, V. L. Potjomkin, and A. Y. Pusep, Measurements of atmospheric condensation nuclei size distributions in Siberia, *J. Aerosol Sci.*, *23*, 191–199, 1992.

Blanchard, D. C., and R. J. Cipriano, Biological regulation of climate, *Nature*, *330*, 526, 1987.

Blanchard, D. C., and A. H. Woodcock, Bubble formation and modification in the sea and its meteorological significance, *Tellus*, *9*, 145–158, 1957.

Browell, E. V., C. F. Butler, S. A. Kooi, M. A. Fenn, R. C. Harriss, and G. L. Gregory, Large–scale variability of ozone and aerosols in the summertime Arctic and sub-Arctic atmosphere, *J. Geophys Res.*, *97*, 16433–16450, 1992.

Clarke, A. D., Atmospheric nuclei in the remote free troposphere, *J. Atmos. Chem.*, *14*, 479–488, 1993.

Clement, C. F., and I. J. Ford, Properties and modeling of global aerosols, *J. Aerosol Sci.*, *28*, (Suppl. 1), 5743–5744, 1997.

Fitzgerald, J. W., Marine aerosols: A review, *Atmos. Environ.*, *25A*, 533–545, 1991.

Heintzenberg, J., Fine particles in the troposphere, a review, *Tellus*, *41B*, 149–160, 1989.

Hofmann, D. J., Twenty years of balloon-borne tropospheric aerosol measurements at Laramie, Wyoming, *J. Geophys. Res.*, *98*, 12753–12766, 1993.

Jaecker-Voirol, A., and P. Mirabel, Heteromolecular nucleation in the sulfuric acid-water system, *Atmos. Environ.*, *23*, 2053–2057, 1989.

Jaenicke, R., Aerosol physics and chemistry, in G. Fischer (Ed.), *Meteorology, Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, Vol. 4, Springer-Verlag, Berlin, 1988, pp. 391–456.

Jaenicke, R., Tropospheric aerosols, in P. V. Hobbs (Ed.), *Aerosol-Cloud-Climate Interactions*, Academic, New York, 1993, pp. 1–31.

Koutsenogii, P. K., N. S. Bufetov, and V. I. Drosdova, Ion composition of atmospheric aerosol near Lake Baikal, *Atmos. Environ.*, *27*, 1629–1633, 1993.

Koutsenogii, P. K., and R. Jaenicke, Number concentration and size distribution of atmospheric aerosol in Siberia, *J. Aerosol Sci.*, *25*, 377–383, 1994.

Leaitch, W. R., and G. A. Isaac, Tropospheric aerosol size distributions from 1982 to 1988 over eastern North America, *Atmos. Environ.*, *25A*, 601–619, 1991.

Monahan, E. C., C. W. Fairall, K. L. Davidson, and P. Jones-Boyle, Observed interrelationships amongst 10m-elevation winds, oceanic whitecaps, and marine aerosols, *Q. J. R. Metereol. Soc.*, *109*, 379–392, 1983.

O'Dowd, C. D., and M. H. Smith, Physicochemical properties of aerosols over the northeast Atlantic: Evidence for wind speed related submicron sea-salt production, *J. Geophys. Res.*, *98*, 1137–1149, 1993.

Prospero, J., Mineral-aerosol transport to the North Atlantic and North Pacific: the impact of African and Asian sources, in A. H. Knap (Ed.), *The Long-Range Atmospheric Transport of*

Natural and Contaminant Substances, NATO ASI series, Kluwer Academic, New York, 1990, pp. 59–82.

Raes, F., R. Van Dingenen, E. Vignati, J. Wilson, J. P. Putaud, J. H. Seinfeld, and P. Adams, Formation and cycling of aerosols in the global troposphere, *Atmos. Environ.*, *34*, 4215–4240, 2000.

Savoie, D. L., and J. M. Prospero, Comparison of oceanic and continental sources of non-sea-salt sulphate over the Pacific Ocean, *Nature*, *339*, 685–687, 1989.

Seinfeld, J. H., and S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, Wiley, New York, 1998.

Wexler, A. S., F. W. Lurmann, and J. H. Seinfeld, Modeling urban and regional aerosols. I. Model development, *Atmos. Environ.*, *28*, 531–546, 1994.

CHAPTER 13

PHOTOCHEMICAL SMOG: OZONE AND ITS PRECURSORS

SANFORD SILLMAN

1 INTRODUCTION

Photochemical smog refers to a number of species that are chemically produced in highly polluted environments in processes driven by sunlight. The most prominent of these species is ozone (O_3), which reaches levels that violate government health standards in urban areas throughout the world. Other components of photochemical smog include peroxyacetyl nitrate (PAN, $CH_3CO_3NO_2$) and nitric and sulfuric acid (HNO_3, H_2SO_4). The latter is associated with the formation of acid aerosols, which have serious impacts on both human health and visibility. Photochemical smog typically forms during conditions characterized by high sunlight (though often with haze and reduced visibility), light winds, and warm temperatures. This chapter focuses on ozone and its precursors.

Episodes with high ozone were first observed in Los Angeles in the 1950s (Haagen-Smit and Fox, 1954) and have generally been found in cities with high automobile traffic. This type of photochemical smog should be distinguished from the type of smog driven by primary emissions (primarily of coal-based SO_2, NO_2, CO, and soot), which characterized the city of London during the early 1900s (Brimblecombe, 1987) and Beijing today. In primary smog, high concentrations are associated with patterns of atmospheric circulation that "trap" the emitted pollutants in an atmospheric layer close to emission sources. The most severe events tend to occur in fall or winter, when atmospheric vertical mixing at ground level is minimal, and may coincide with fog (hence the origin of the word *smog* for the combination of smoke and fog). By contrast, photochemical smog can only occur in meteorological conditions that favor photochemical activity (i.e., high sunlight, warm temperatures)

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

and not necessarily during conditions with restricted meteorological dispersion. The most severe events have occurred in large urban areas with warm, dry climates (Los Angeles, Mexico City, Athens). However severe photochemical smog has been observed in virtually all major cities of North America and Europe and more recently in developing nations. Although the most severe episodes have occurred in locations with high automobile traffic, elevated O_3 has also been associated with coal-fired power plants (Miller et al., 1978; White et al., 1983; Gillani and Pleim, 1996; Ryerson et al., 2001). In the eastern United States and in Europe elevated O_3 occurs in regionwide events and is characterized by transport over distances of 500 km or more. Elevated O_3 has also been found in association with biomass burning in the tropics (see Chapter 14).

The relation between ozone formation and precursor emissions has been the subject of much uncertainty and controversy. Ozone is formed from two general classes of precursors: hydrocarbons (including oxygenated organic species)* and nitrogen oxides ($NO + NO_2$, or NO_x). The chemistry of ozone formation typically falls into one of two recognizable patterns: a NO_x-limited regime in which the rate of formation increases with NO_x and is largely independent of hydrocarbon concentrations and a hydrocarbon-limited (or light-limited) regime in which the rate of formation increases with hydrocarbons and decreases with increasing NO_x. An analogous split into NO_x-limited and light-limited regimes also occurs in the remote troposphere. The split into NO_x-limited and HC-limited regimes has generated a debate on policy, especially in the United States, concerning the best way to reduce urban ozone. Because this represents a major uncertainty associated with ozone formation, much of this chapter will address the complex relation between ozone, NO_x, and HC.

2 GENERAL FEATURES OF PHOTOCHEMICAL SMOG

Diurnal and Seasonal Cycle

Ozone and other secondary reaction products show a pronounced diurnal cycle with peak concentrations typically occurring in late afternoon. The diurnal cycle of O_3 shows a sharp contrast with the diurnal cycle of primary species, including NO_x, HC, and CO (see Fig. 1). The primary species typically have peak concentrations in early morning and much lower concentrations during the daytime as concentrations are diluted through the process convection-driven vertical mixing. Because production of O_3 requires sunlight, peak concentrations often occur at the time of maximum vertical mixing and often coincide with diurnal minima in precursor concentrations. The diurnal cycle of O_3 is also influenced by nighttime removal of O_3 near the ground (through surface deposition or through reaction with directly emitted NO). Especially in urban areas, O_3 concentrations near the surface are often very low at night. The characteristic increase in O_3 during the morning hours (6 to 10 A.M.) is usually driven by convective mixing that breaks up the nighttime inversion at the

*The family consisting of hydrocarbons and oxygenates such as formaldehyde, HCHO, is properly referred to by acronyms such as volatile organic compounds (VOC) or reactive organic gases (ROG). In this chapter they will be referred to collectively as hydrocarbons (HC).

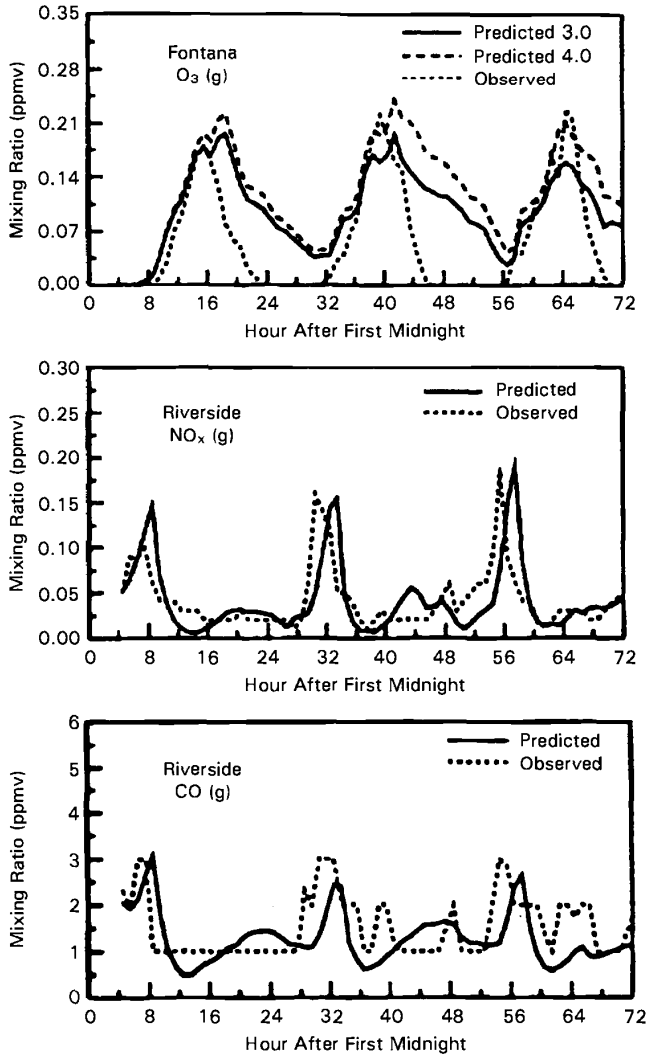


Figure 1 Time series for O_3, CO, NO, and NO_2 (in parts per million, ppmv) vs. hour at Riverside, CA, August 26–28, 1988. Dashed lines represent measurements, and solid lines represent model predictions. From Jacobson et al. (1996).

surface and mixes down air (from 100 to 1000 m above the surface) with higher O_3. The subsequent rise in O_3 after 10 A.M. is often associated with photochemical production.

The city of Los Angeles is especially susceptible to high-ozone events because it frequently sees a combination of high sunlight, warm temperatures, and a low-level thermal inversion (typically 500 to 1000 m above the surface) during the daytime. In most other cities the conditions that favor ozone formation (sunlight and warm temperatures) coincide with vigorous vertical mixing (up to 2000 m), which has a moderating effect on ozone concentrations. Thermal inversions, which trap pollutants near the ground, are more commonly associated with cold temperatures and often with fog. These conditions would produce high concentrations of primary pollutants but not ozone.

Unless stated otherwise, the discussion of ozone concentrations below refers to the diurnal peak or near-peak concentrations that occur during the afternoon.

Concentrations and Regional Transport

The global background concentration of O_3 near the surface is 20 to 40 ppb parts per billion (ppb), although these values probably represent an increase in comparison with preindustrial concentrations. There have also been episodes in which high concentrations of O_3 originating in the upper troposphere, ultimately of stratospheric origin, may have been transported to the surface.

During air pollution events in the United States and Europe, peak O_3 frequently exceeds 125 ppb, which is the current government health standard in the United States.* In Los Angeles during the 1970s and 1980s air quality violations (i.e., $O_3 < 125$ ppb) were reported on approximately 180 days per year. In the 1990s the frequency of violation has been lowered to 90 days per year. Most other major cities in the United States record violations on 5 to 10 days per year. In Europe, ozone exceeds 125 ppb on just a few days per year, while in Mexico City at present, ozone exceeds 125 ppb on 200 days per year. Concentrations above 200 ppb are found only during the most severe events, and concentrations as high as 490 ppb have been observed in Los Angeles and in Mexico City.

In addition, 80 to 100 ppb ozone is frequently observed in rural areas of the eastern United States and Europe during regional events. In these events, air with ozone concentrations above 80 ppb frequently extends over a 1000×1000 km region and extends vertically to 1000 to 2000 m above the surface (e.g., Clarke and Ching, 1983). These events are often associated with stagnant high pressure systems in which air may be trapped under a subsidence inversion at ~2000 m. An example in eastern North America is shown in Figure 2. In the example, ozone above 120 ppb

*As of 1997, a 1-h average concentration in excess of 125 ppb constitutes an air quality violation in the U.S. metropolitan areas that record violations of the 1-h standard on more than 3 days over a 3-year period are held in violation of clean air laws and are asked to submit a plan for pollution reduction. Since 1977, most U.S. cities have been continually in violation. It was recently proposed that the 1-h air quality standard be replaced with a standard based on 8-hour average concentrations, in which an 8-h average concentration in excess of 85 ppb is counted as an air quality violation.

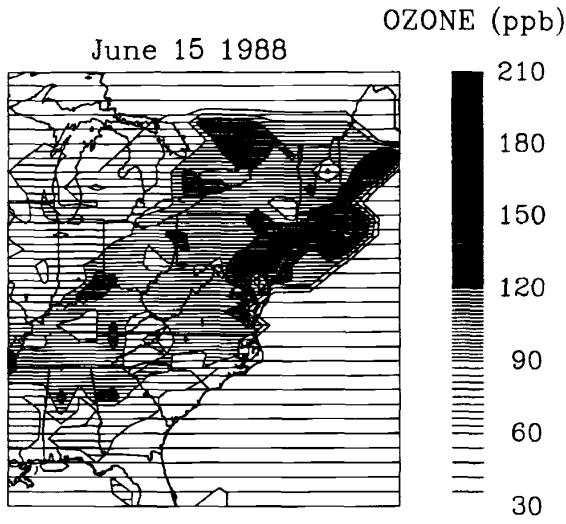


Figure 2 Peak ozone concentrations in the eastern United States during a severe air pollution event (June 15, 1988) based on surface observations at 350 EPA monitoring sites. The shadings represent values of 30 to 60 ppb (lightest shading) to 180 to 210 ppb (darkest shading) with 30-ppb intervals in between. Values reported for Canada and the Atlantic Ocean are inaccurate since no observations were available for these locations. First printed in Sillman (1993).

was found in many metropolitan areas, especially in the corridor extending from Washington to New York and Boston. However, ozone above 90 ppb covered a much larger area extending from Kentucky to Maine. Although ozone concentrations above 120 ppb were generally associated with plumes from specific urban areas (or from coal-fired power plants), concentrations of 90 to 100 ppb were found at rural sites throughout the region. In addition, unusually high ozone (> 200 ppb) was found in Acadia National Park in Maine. The high ozone in Maine is most likely due to transport from Boston (300 km distant) and the New York area (700 km distant).

Environmental and Health Impacts

The impact of ozone and acid aerosols on human health has been the subject of intense scrutiny. Ozone and aerosols have been associated with a variety of lung ailments. Short-term symptoms (including lung inflammation, asthmatic responses, and measured impairment of lung functions) have been found in experiments in response to ozone concentrations as low as 120 ppb. High-ozone events have been correlated with increased admissions to hospitals for respiratory diseases and with increased mortality rates. For a summary of findings, see Lippman (1993) and Bascomb et al. (1996).

In addition, ozone concentrations of 80 ppb have been found to cause damage both to forests and to agricultural crops. Crop damage from ozone in the United States has been estimated to cause monetary losses of $1 to 2 billion per year. For a summary of findings, see U.S. Congress (1989) and National Research Council (NRC, 1991).

Dependence on Temperature

As stated above, ozone in polluted regions shows a strong dependence on temperature. This dependence on temperature is important as a basis for understanding variations in ozone concentrations from year to year or between cities. As shown in Figure 3, elevated ozone is always associated with temperatures in excess of 20°C and is often with temperatures above 30°C. In the eastern United States and Europe, year-to-year variations in ozone concentrations are often the result of variations in temperature and cloud cover, rather than in changes in emission of pollutants.

The reason for the dependence on temperature is due largely to the chemistry of ozone formation. Cardelino and Chameides, (1990) and Sillman and Samson (1995) found that the temperature dependence was associated with the temperature-dependent decomposition rate of PAN. PAN becomes longer lived at lower temperatures, and formation of PAN results in the removal of NO_x, hydrocarbons, and odd hydrogen radicals (described below), all of which suppress ozone formation. PAN, also a component of photochemical smog, tends to reach maximum values at intermediate temperatures (5 to 10°C). Jacob et al. (1993) proposed that ozone correlates with temperature partly because the meteorological conditions that favor ozone formation (high solar radiation and light winds) tend to be associated with warm temperatures. In addition, the emission rate of biogenic hydrocarbons (a major ozone precursor, discussed below) increase sharply with increasing temperature. Ozone is affected by temperature only in polluted regions. Temperature apparently has little impact on ozone production at the global scale (Sillman and Samson, 1995).

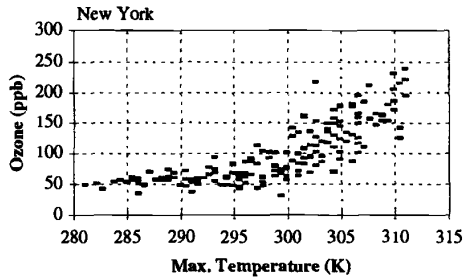


Figure 3 Diurnal peak O_3 (ppb) vs. maximum surface temperature observed in the New York–New Jersey–Connecticut metropolitan area for April 1 through September 30, 1988. From Sillman and Samson, 1995.

Role of Biogenic Hydrocarbons

The main precursors of photochemical smog, NO_x and hydrocarbons, are emitted into the atmosphere by a variety of human activities—transport (chiefly automobiles), coal-fired industry (especially electric power plants), and biomass burning. However, significant amounts of hydrocarbons occur naturally and are emitted by vegetation, primarily from trees. The most important of these biogenic hydrocarbons are isoprenes (C_5H_8), emitted by oaks and other deciduous trees, and α- and β-pinenes ($C_{10}H_{16}$), which are emitted from conifers. These species react chemically in the same way as anthropogenic hydrocarbons and can function as precursors to photochemical smog. In the United States it is estimated that emission of biogenic hydrocarbons equals or exceeds emission of anthropogenic hydrocarbons (Geron et al., 1994). Even in urban areas biogenic hydrocarbons can account for a significant fraction of total hydrocarbon emissions and can have a large impact on the formation of smog (Chameides et al., 1988). The impact of isoprene is especially large because it reacts rapidly, with a chemical lifetime of one hour or less. Consequently even small amounts of isoprene (0.5 ppb) can have a large impact on ozone.

It should be emphasized that naturally occurring hydrocarbons will not lead to the formation of photochemical smog in the absence of human activities because smog formation requires NO_x in addition to hydrocarbons. Although some NO_x is emitted naturally through biological activity, naturally occurring NO_x emissions are too small to allow significant formation of O_3 and other components of smog. Biogenic NO_x is estimated to be 7% of total NO_x emissions in the United States (Williams et al., 1992) and most of this is associated with agriculture (especially with the use of nitrate fertilizer).

Ozone Production Efficiency

The ozone production efficiency represents the rate of production of ozone divided by the loss rate for NO_x [$P(O_3)/L(NO_x)$]. Liu et al. (1987) first introduced the concept of ozone production efficiency and used it as a basis for estimating global production of ozone as a function of estimated NO_x emissions. A central feature of the ozone production efficiency is the tendency toward lower values in more polluted environments. Recent estimates suggest that ozone production efficiency is 10 to 30 in the remote troposphere but just 3 to 5 in urban areas.

3 RELATION BETWEEN OZONE, NO_x, AND HYDROCARBONS

The relation between ozone, NO_x and hydrocarbons can be illustrated by an isopleth plot (Fig. 4), which shows instantaneous rates of ozone production as a function of NO_x and hydrocarbon concentrations. It can be seen that ozone production is a highly nonlinear process, especially with regard to NO_x. Ozone production as a function of NO_x shows well-defined local maxima, usually at a specific HC/NO_x ratio. This region of maximum ozone (the "ridge line") can be thought of as a divide

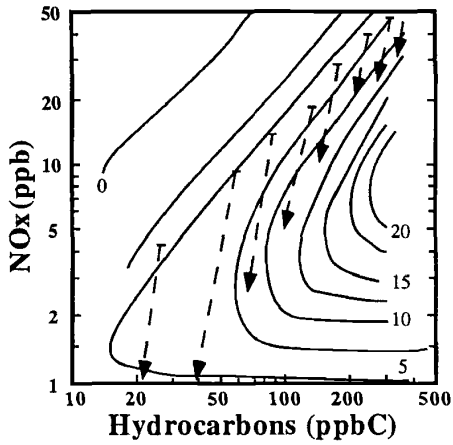


Figure 4 Isopleths giving net rate of ozone production (ppb per hour, daytime average, solid line) as a function of ROG (ppbC) and NO_x (ppb). The dashed lines and arrows show the calculated evolution of ROG and NO_x concentrations in a series of air parcels over an 8-h period (9 A.M. to 5 P.M.), each with initial $ROG/NO_x = 6$ and speciation typical of urban centers in the United States, based on calculations shown in Milford et al. (1994).

between two regimes with different photochemical behavior. Above the ridge line (with low HC/NO_x ratios), ozone production rates increase with increasing HC but decrease with increasing NO_x (hydrocarbon-limited regime). Below the ridge line (with high HC/NO_x ratios) ozone production rates increase with increasing NO_x and will be largely unaffected by changes in hydrocarbons (NO_x-limited regime). The existence of these two regimes has an enormous impact on public policy because it affects the choice of control strategies for reducing high ozone levels. If ozone production is dominated by NO_x-limited chemistry, then reductions in NO_x emissions would be necessary to reduce ozone concentrations. If production is dominated by HC-limited chemistry, then reductions in hydrocarbons would be needed. There is also a complex relation between NO_x, HC, and particulates, which also affects policy choices (Meng et al., 1997).

An important feature of HC–NO_x chemistry is the tendency for polluted air to evolve toward the NO_x-limited regime as the air mass ages and moves downwind. Air is most likely to show HC-limited chemistry when it is close to emission sources, especially in large cities. As the air mass ages, the HC/NO_x ratio increases and the chemistry shifts to the NO_x-limited regime. This shift from HC-limited to NO_x-limited chemistry as polluted air ages is illustrated by the air parcel trajectories in Figure 4.

In terms of photochemical mechanisms, the split into NO_x-limited and HC-limited regimes is closely related to the chemistry of odd hydrogen, defined as the sum of OH, HO_2, and RO_2 radicals (where R represents a hydrocarbon chain). The sequence of reactions that lead to photochemical smog (including production of both

ozone and acid aerosols) is usually initiated by reactions that involve the OH radical, and availability of OH (which is produced from reactions involving sunlight, O_3 and H_2O) often controls the rate of ozone production. The split into NO_x- and HC-limited regimes is determined by the loss mechanism for odd hydrogen (Sillman, 1995; Sillman et al., 1990; Kleinman 1991, 1994; summarized in Sillman, 1999). The reaction sequences are shown in the postscript to this chapter.

Kleinman (1991, 1994) has shown that the split between NO_x- and HC-limited regimes can be explained simply in terms of the supply of NO_x relative to the source of odd hydrogen. NO_x-limited chemistry occurs when the supply of odd hydrogen exceeds the supply of NO_x, while HC-limited chemistry (also referred to as light-limited chemistry) occurs when the supply of NO_x exceeds that of odd hydrogen. This explanation is useful because it provides a conceptual basis for understanding how HC–NO_x chemistry varies from location to location and from event to event. For example, HC-limited chemistry is most likely to occur in large cities and during severe events, when the supply of NO_x is largest, and also during periods of low sunlight, which limits the source of odd hydrogen. NO_x-sensitive chemistry is more likely in smaller cities and during more moderate events (i.e., with lower NO_x) and in far downwind locations (Milford et al., 1994). These trends in HC–NO_x chemistry are all consistent with Kleinman's description.

The following is a summary of factors that affect the variation between HC-limited and NO_x-limited chemistry.

HC:NO_x Ratio

As illustrated in Figure 4, HC-sensitive chemistry is associated with low HC/NO_x ratios and NO_x-sensitive chemistry is associated with high HC/NO_x. The importance of HC/NO_x ratios was identified in the early research into the causes of photochemical smog in the 1950s (Haagen-Smit and Fox, 1954).

For many years, the U.S. Environmental Protection Agency (EPA) used a rule of thumb that HC-NO_x chemistry could be deduced based on the HC/NO_x ratio at 6 to 9 A.M., where a ratio of 10* or less was presumed to correspond with HC-sensitive chemistry and a ratio of 20 or more would correspond with NO_x-sensitive chemistry. This approach has been discredited (e.g., NRC, 1991) because it failed to take into account many of the other factors that affect HC–NO_x chemistry, described below. The chemical impact of hydrocarbons also depends on the reactivity of the hydrocarbon species, so that NO_x-sensitive chemistry is more likely when HC reactivity is higher. It is often useful to think of hydrocarbons and HC/NO_x ratios in terms of reactivity-weighted sums rather than just total concentration.

*Hydrocarbon concentrations are often expressed in parts per billion carbon (ppbC), which represents a sum of species concentrations in ppb weighted by the number of carbon atoms contained. HC/NO_x ratios are expressed in pppC/ppb.

Biogenic Hydrocarbons

The inclusion of biogenic hydrocarbons in analyses of photochemical smog often has a large impact on HC–NO_x chemistry. Biogenic hydrocarbons cause an increase in HC/NO_x ratios and therefore cause a shift toward NO_x-sensitive chemistry. The impact of biogenic hydrocarbons is often overlooked because (i) biogenic hydrocarbons are extremely reactive, and consequently have an impact on chemistry out of proportion to their ambient concentrations; and (ii) biogenic emissions are zero at night and low during the morning hours, and are therefore underrepresented in the traditional morning HC/NO_x ratio.

Historically, the role of biogenic hydrocarbons on urban ozone formation was not recognized until 1988. More recently, it has been shown that emission estimates used by the U.S. EPA underestimated biogenic emissions by factors of 3 or more (Geron et al., 1994). Unpublished results from model calculations suggest that use of the higher biogenic emission estimates would cause a shift from primarily HC-sensitive chemistry to primarily NO_x-sensitive chemistry in many cities of the eastern United States. The impact of biogenic hydrocarbons is smaller in Europe (Simpson, 1995).

Geographical Variation

The pattern of geographical variation in HC–NO_x chemistry is largely associated with the photochemical aging process. As stated above, fresh emissions are often in an HC-limited state but evolve toward NO_x-limited chemistry as the air mass ages. In addition, the total accumulation of ozone in a fully aged air mass appears to be controlled entirely by NO_x rather than by HC. In other words, a reduction in hydrocarbons in an HC-limited region has the effect of deferring ozone production until an air mass moves downwind and disperses, but may have little effect on the number of ozone molecules that is produced once the chemistry has run to completion.

The contrast between HC-limited chemistry in an urban center and NO_x-limited chemistry in downwind suburban regions has been dominated most extensively for Los Angeles (Milford et al., 1989, 1994). Extensive measurements and model calculations have supported the view that downtown Los Angeles has HC-limited chemistry. By contrast, ozone in rural locations in the eastern United States (representing photochemically aged air) is usually sensitive to NO_x rather than HC, although there are exceptions. The highest ozone concentrations typically occur in suburban locations approximately 6 h downwind of major urban centers. These locations represent an intermediate situation between the HC-limited chemistry of urban centers and NO_x-limited chemistry in far downwind locations. It is often uncertain whether peak ozone concentrations are associated with HC-limited or NO_x-limited chemistry, and predictions (derived from model calculations) are frequently dependent on model assumptions, e.g., about emission rates, winds, and vertical mixing.

It should be emphasized that predictions for HC–NO_x sensitivity for individual locations are all highly uncertain at this time. HC–NO_x predictions are often based on model calculations with little supporting evidence from ambient measurements.

The most detailed analyses, including both model calculations and analyses of ambient measurements, have been done for rural sites in the eastern United States (e.g., Buhr et al., 1995; Roselle and Schere, 1995; Jacob et al., 1995; Sillman, 1995; Trainer et al., 1993) and for Los Angeles (e.g., Milford et al., 1989; Jacobson et al., 1996) Other area evaluations have been done for Atlanta (Sillman et al., 1995), Mexico City (Sosa et al., 2000), and Europe (Simpson, 1995). For a more complete summary, see Sillman, (1999) and NRC (1991).

Evaluation of Ozone Production through Measurements

Two types of ambient measurements are especially important for evaluating the impact of photochemistry as a source of O_3. One is the correlation between O_3 and CO (e.g., Parrish et al., 1993). Because CO is primarily a product of human activities (either industry or biomass burning), a positive correlation between these species is interpreted as a signal for photochemical smog, especially in the remote troposphere. A second measurement is the correlation between O_3 and the sum of total reactive nitrogen (NO_y, including NO_x, PAN, HNO_3, and other organic nitrates) and between O_3 and the sum of NO_x reaction products (NO_y–NO_x, or NO_z). Because O_3 and NO_z are both produced by similar photochemical processes, there is a strong correlation between these species in polluted environments at times of photochemical activity.

The correlation between O_3 and NO_z is also interpreted as a measure of ozone production efficiency, defined as the ratio of net production of ozone to the loss rate for NO_x [$P(O_3)/L(NO_x)$]. The slope between O_3 and NO_z is determined partly by the ozone production efficiency but is also influenced by atmospheric removal processes, especially for HNO_3. The ozone production efficiency is often used as a basis for interpreting ozone chemistry (e.g., Liu et al., 1987). In addition, Sillman (1995, 1998, 1999) has proposed that the value of the ratio O_3/NO_z can be used as an "indicator" for NO_x-sensitive versus HC-sensitive ozone chemistry.

4 CHEMISTRY OF OZONE FORMATION

Ozone is produced by a reaction sequence that is initiated by reaction of hydrocarbons or CO with the OH radical. Although individual hydrocarbons follow complex reaction pathways, they often conform to the following pattern:

$$HC + OH \xrightarrow{[O_2]} RO_2 + H_2O \quad (1)$$

$$CO + OH \xrightarrow{[O_2]} HO_2 + CO_2 \quad (2)$$

RO_2 represents a hydrocarbon chain with O_2 attached. The group of RO_2 radicals and HO_2 all react rapidly with NO:

$$RO_2 + NO \xrightarrow{[O_2]} R'CHO + HO_2 + NO_2 \quad (3)$$

$$HO_2 + NO \xrightarrow{[O_2]} OH + NO_2 \quad (4)$$

resulting in an intermediate hydrocarbon by-product ($R'CHO$) and NO_2. This is followed by photolysis of NO_2:

$$NO_2 + h\nu \rightarrow NO + O \quad (5)$$

The resulting oxygen atom rapidly combines with O_2 to form ozone ($O + O_2 + M \rightarrow O_3 + M$).

The rate of formation of ozone and other smog elements, including sulfate and nitrate aerosols, depends critically on the OH radical, which initiates the reaction sequence. The complex dependence of ozone on NO_x and hydrocarbons is closely linked to the chemistry of OH and associated radical species, including HO_2 and RO_2 radicals. Because the reaction sequence (1) through (4) operate on the radicals OH, HO_2, and RO_2 without changing the sum $OH + HO_2 + RO_2$, it is useful to regard the latter sum as a family of species (odd hydrogen). Much of the complexity of ozone chemistry can be understood by analyzing sources and sinks for this family. Odd hydrogen sources are almost all photolytic reactions and include the following:

$$O_3 + h\nu \xrightarrow{[H_2O]} 2OH \quad (6)$$

$$HCHO + h\nu \xrightarrow{[O_2]} 2HO_2 + CO \quad (7)$$

Odd hydrogen is removed by reactions that produce hydrogen peroxides and nitric acid:

$$HO_2 + HO_2 \rightarrow H_2O_2 + O_2 \quad (8)$$

$$RO_2 + HO_2 \rightarrow ROOH + O_2 \quad (9)$$

$$OH + NO_2 \rightarrow HNO_3 \quad (10)$$

Formation of peroxyacetyl nitrate (PAN) is also a significant sink for odd hydrogen.

It is possible to derive an analytic solution for OH and for the rate of production of ozone as a function of NO_x and HC based on the above reactions (Sillman et al., 1990, 1995). The solution has the form of a fourth degree polynomial for OH, NO_x, and HC (or for ozone production, NO_x, and HC) and reproduces many of the qualitative features of OH and ozone production as a function of NO_x and HC (Fig. 4). HC-limited chemistry occurs when formation of nitric acid (10) represents the dominant loss mechanism for odd hydrogen. In this situation reactions (6), (7), and (10) form an approximate steady state that determines OH. Increased NO_x

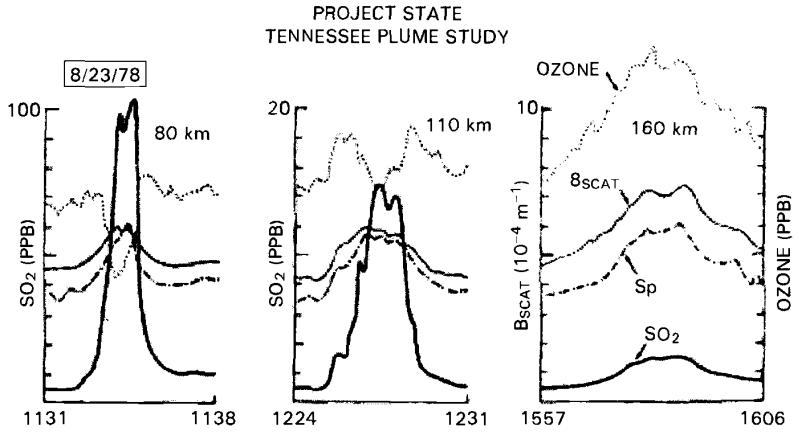


Figure 5 Stages in the chemical development of a power plant plume. The three sets of profiles show measurements of SO_2 (surrogate for NO_x, heavy solid line), ozone (dotted line), particulate sulfur (S_p, line-dot-line), all in ppb; and the light-scattering coefficient (B_{scat}, 10^{-4}/m, light solid line) made during crosswind aircraft traverses through the plume of the Cumberland power plant in NW Tennessee on August 23, 1978. The traverses at 80, 110, and 160 km downwind distances illustrate the "early," the "intermediate," and the "mature" stages of chemical development of the plume, respectively. From Gillani et al., 1996.

causes a decrease in OH and a consequent decrease in the rate of ozone production [approximately equal to the rate of (1)]. Increased HC causes a modest increase in OH [due to (7)] and a larger increase in the rate of (1), which leads to increased ozone production. NO_x-limited chemistry occurs when formation of peroxides [(8) and (9)] represents the dominant sink for odd hydrogen. In this situation the sum $HO_2 + RO_2$ is determined by the steady state between (6), (8), and (9) and is relatively insensitive to changes in NO_x or HC. The rate of ozone formation, approximately equal to the rate of reactions (3) and (4), increases with increasing NO_x and is largely unaffected by HC.

At nighttime O_3 is removed by reaction with NO, as follows:

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{11}$$

During the daytime reactions (5) and (11) both occur rapidly, but the combination has little effect on ozone concentrations. However, at nighttime, (5) does not occur and (12) results in removal of O_3. Reaction (11) also causes a decrease in ozone during the daytime in the vicinity of a large emission source of NO, e.g., coal-fired power plants. Power plant plumes typically show a decrease in O_3 immediately downwind of the plant, followed by recovery and subsequent increase in O_3 as the ozone-forming reactions (1) to (4) occur (see Fig. 5) (White et al., 1983; Gillani and Pleim, 1996). This pattern of reduced O_3 near the plume source followed by

enhanced O_3 downwind is similar to the pattern of evolution of urban plumes with HC-limited chemistry near emission sources and NO_x-limited chemistry further downwind.

ACKNOWLEDGMENTS

Support for this work was provided by the U.S. National Science Foundation (grant #ATM-9713567).

REFERENCES

Bascomb, R., P. A. Bromberg, D. L. Costa, R. Devlin, D. W. Dockery, M. W. Frampton, W. Lambert, J. M. Samet, F. E. Speizer, and M. Utell. Health effects of outdoor air pollution. *Am. J. Resp. Crit. Care Med.*, *153*, 477–498, 1996.

Brimblecombe, P., *The Big Smoke: A History of Air Pollution in London since Medieval Times*, Methuen, London, 1987.

Buhr, M., D Parrish, J. Elliot, J. Holloway, J. Carpenter, P. Goldan, W. Kuster, M. Trainer, S. Montzka, S. McKeen, and F. C. Fehsenfeld, Evaluation of ozone precursor source types using principal component analysis of ambient air measurements in rural Alabama. *J. Geophys. Res.*, *100*, 22853–22860, 1995.

Cardelino, C. A., and W. L. Chameides, Natural hydrocarbons, urbanization, and urban ozone, *J. Geophys. Res.*, *95*, 13971–13979, 1990.

Chameides, W. L., R. W. Lindsay, J. Richardson, and C. S. Kiang, The role of biogenic hydrocarbons in urban photochemical smog: Atlanta as a case study, *Science*, 241, 1473–1474, 1988.

Clarke, J. F., and J. K. S. Ching, Aircraft observations of regional transport of ozone in the northeastern United States, *Atmos. Environ.*, *17*, 1703–1712, 1983.

Geron, C. D., A. B. Guenther, and T. E. Pierce, An improved model for estimating emissions of volatile organic compounds from forests in the eastern United States, *J. Geophys. Res.*, *99*, 12773–12791, 1994.

Gillani, N. V., and J. E. Pleim, Sub-grid-scale features of anthropogenic emissions of NO_x and VOC in the context of regional Eulerian models, *Atmos. Environ.*, *30*, 2043–2059, 1996.

Haagen-Smit, A. J., and M. M. Fox, Photochemical ozone formation with hydrocarbons and automobile exhaust, *J. Air Pollut. Control Assoc. 4*, 105–109, 1954.

Jacob, D. J., B. G. Heikes, R. R. Dickerson, R. S. Artz, and W. C. Keene, Evidence for a seasonal transition from NO_x- to hydrocarbon-limited ozone production at Shenandoah National Park, Virginia, *J. Geophys. Res.*, *100*, 9315–9324, 1995.

Jacob, D. J., J. A. Logan, G. M. Gardner, R. M. Yevich, C. M. Spivakowsky, S. C. Wofsy, S. Sillman, and M. J. Prather, Factors regulating ozone over the United States and its export to the global atmosphere, *J. Geophys. Res.*, *98*, 14817–14827, 1993.

Jacobson, M. Z., R. Lu, R. P. Turco, and O. P. Toon, Development and application of a new air pollution modeling system—Part I: Gas-phase simulations. *Atmos. Environ.*, *30*, 1939–1963, 1996.

Kleinman, L. I., Seasonal dependence of boundary layer peroxide concentration: The low and high NO_x regimes, *J. Geophys. Res.*, *96*, 20721–20734, 1991.

Kleinman, L. I., Low and high-NO_x tropospheric photochemistry, *J. Geophys. Res.*, *99*, 16831–16838, 1994.

Lippman, M., Health effects of tropospheric ozone: Review of recent research findings and their implications to ambient air quality standards, *J. Expos. Anal. Environ. Epidemiol.*, *3*, 103–128, 1993.

Liu, S. C., M. Trainer, F. C. Fehsenfeld, D. D. Parrish, E. J. Williams, D. W. Fahey, G. Hubler, and P. C. Murphy, Ozone production in the rural troposphere and the implications for regional and global ozone distributions, *J. Geophys. Res.*, *92*, 4191–4207, 1987.

Meng, Z., D. Dabdub, and J. H. Seinfeld, Chemical coupling between atmospheric ozone and particulate matter, *Science*, *277*, 116–119, 1997.

Milford, J., D. Gao, S. Sillman, P. Blossey, and A. G. Russell, Total reactive nitrogen (NO_y) as an indicator for the sensitivity of ozone to NO_x and hydrocarbons, *J. Geophys. Res.*, *99*, 3533–3542, 1994.

Milford, J., A. G. Russell, and G. J. McRae, A new approach to photochemical pollution control: Implications of spatial patterns in pollutant responses to reductions in nitrogen oxides and reactive organic gas emissions, *Environ. Sci. Technol.*, *23*, 1290–1301, 1989.

Miller, D. F., A. J. Alkezweeny, J. M. Hales, and R. N. Lee, Ozone formation related to power plant emissions, *Science*, *202*, 1186–1188, 1978.

National Research Council (NRC), Committee on Tropospheric Ozone Formation and Measurement, *Rethinking the Ozone Problem in Urban and Regional Air Pollution*, National Academy Press, Washington, DC, 1991.

Parrish, D. D., J. S. Holloway, M. Trainer, P. C. Murphy, G. L. Forbes, and F. C. Fehsenfeld, Export of North American ozone pollution to the North Atlantic Ocean, *Science*, *259*, 1436–1439, 1993.

Roselle, S. J., and K. L. Schere. Modeled response of photochemical oxidants to systematic reductions in anthropogenic volatile organic compound and NO_x emissions, *J. Geophys. Res.*, *100*, 22929–22941, 1995.

Ryerson, T. B., M. Trainer, J. S. Holloway, D. D. Parrish, L. G. Huey, D. T. Sueper, G. J. Frost, S. G. Donnelly, S. Schauffler, E. L. Atlas, W. C. Kustler, P. D. Goldman, G. Hubler, J. F. Meagher, and F. C Frehsenfeld, Observations of ozone formation in power plant plumes and implications for ozone control stategies, *Science*, **292**, 719–723, 2001.

Sillman, S. Tropospheric ozone: The debate over control strategies, *Annu. Rev. Energy Environ.*, *18*, 31–56, 1993.

Sillman, S., The use of NO_y, H_2O_2 and HNO_3 as indicators for O_3-NO_x-ROG sensitivity in urban locations, *J. Geophys. Res.*, *100*, 14175–14188, 1995.

Sillman, S., The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments. Millenial review series, *Atmos. Environ.*, *33*(12), 1821–1845, 1999.

Sillman, S., K. Al-Wali, F. J. Marsik, P. Nowatski, P. J. Samson, M. O. Rodgers, L. J. Garland, J. E. Martinez, C. Stoneking, R. E. Imhoff, J-H. Lee, J. B. Weinstein-Lloyd, L. Newman, and V. Aneja, Photochemistry of ozone formation in Atlanta, GA: Models and measurements, *Atmos. Environ.*, *29*, 3055–3066, 1995.

Sillman, S., D. He, M. Pippin, P. Daum, L. Kleinman, J. H. Lee and J. Weinstein-Lloyd, Model correlations for ozone, reactive nitrogen and peroxides for Nashville in comparison with

measurements: Implications for VOC-NO_x sensitivity, *J. Geophys. Res.*, *103*, 22629–22644, 1998.

Sillman, S., J. A. Logan, and S. C. Wofsy, The sensitivity of ozone to nitrogen oxides and hydrocarbons in regional ozone episodes, *J. Geophys. Res.*, *95*, 1837–1851, 1990.

Sillman, S. and P. J. Samson, The impact of temperature on oxidant formation in urban, polluted rural and remote environments, *J. Geophys. Res.*, *100*, 11497–11508, 1995.

Simpson, D., Biogenic emissions in Europe, 2, Implications for ozone control strategies, *J. Geophys. Res.*, *100*, 22891–22906, 1995.

Sosa, G., J. West, F. San Martini, L. T. Molina and M. J. Molina, "Air Quality Modeling and Data Analysis for Ozone and Particulates in Mexico City." MIT Integrated Program on Urban, Regional and Global Air Pollution Report No. 15, 76 pages, October 2000, available from http://eaps.mit.edu/megacities/index.html.

Trainer, M., D. D. Parrish, M. P. Buhr, R. B. Norton, F. C. Fehsenfeld, K. G. Anlauf, J. W. Bottenheim, Y. Z. Tang, H. A. Wiebe, J. M. Roberts, R. L. Tanner, L. Newman, V. C. Bowersox, J. M. Maugher, K. J. Olszyna, M. O. Rodgers, T. Wang, H. Berresheim, and K. Demerjian, Correlation of ozone with NO_y in photochemically aged air, *J. Geophys. Res.*, *98*, 2917–2926, 1993.

U.S. Congress, Office of Technology Assessment, *Catching Our Breath: Next Steps for Reducing Urban Ozone*, OTA-O-412, U.S. Government Printing Office, Washington, DC, 1989.

White, W. H., D. E. Patterson, and W. E. Wilson, Jr., Urban exports to the nonurban troposphere: Results from project MISTT, *J. Geophys. Res.*, *88*, 10745–10752, 1983.

Williams, E. J., A. Guenther, and F. C. Fehsenfeld, An inventory of nitric oxide emissions from soils in the United States, *J. Geophys. Res. 97*, 7511–7519, 1992.

CHAPTER 14

BIOMASS BURNING

ANNE M. THOMPSON

1 INTRODUCTION

Biomass fires are both natural and anthropogenic in origin. The natural trigger is lightning, which leads to mid- and high-latitude fires and episodes of smoke and pollution associated with them. Lightning is also prominent in tropical regions when the dry season gives way to the wet season and lightning in convective systems ignites dry vegetation.

Atmospheric consequences of biomass fires are complex. When considering the impacts of fires for a given ecosystem, inputs of fires must be compared to other processes that emit trace gases and particles into the atmosphere. Other processes include industrial activity, fires for household purposes, and biogenic sources, which may themselves interact with fires. That is, fires may promote or restrict biogenic processes (Fig. 1).

Several books have presented various aspects of fire interactions with atmospheric chemistry (Levine, 1991, 1996; Crutzen and Goldammer, 1993) and a cross-disciplinary review of a 1992 fire-oriented experiment appears in *SAFARI: The Role of Southern African Fires in Atmospheric and Ecological Environments* (van Wilgen et al., 1997). The IGAC/BIBEX core activity (see acronyms at end of chapter) has sponsored field campaigns that integrate multiple aspects of fires—ground-based measurements with an ecological perspective, atmospheric measurements with chemical and meteorological components, and remote sensing (Table 1).

This chapter presents two aspects of biomass fires and the environment. Namely, the relationship between biomass burning and ozone is described, starting with a brief description of the chemical reactions involved and illustrative measurements and interpretation. Second, because of the need to observe biomass burning and its consequences globally, a summary of remote-sensing approaches to the study of fires

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

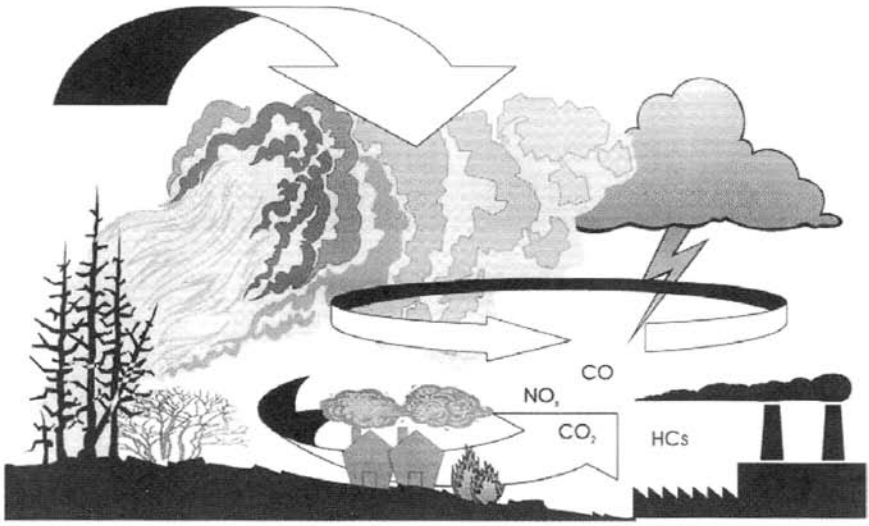


Figure 1 Schematic of processes in tropics with significant production of trace gases—CO, hydrocarbons, or NO—that contribute to tropospheric ozone formation. Biomass fires are major sources of CO, hydrocarbons, and NO, but lightning and soils contribute to NO in the upper troposphere and boundary layer, respectively. Soils release CO as well, under certain conditions and vegetative production is a large source of hydrocarbons. Isoprene from vegetation is oxidized to form CO and more highly reactive oxygenated hydrocarbon intermediates. Besides burning of biomass, burning of wood for fuel use and industrial combustion release the ozone precursors, CO, hydrocarbons, and NO.

and trace gases is given. Examples in this chapter are restricted to tropical burning for matters of brevity and because most burning activity globally is within this zone.

2 CHEMICAL REACTIONS: OZONE FORMATION AND EFFECTS OF FIRES ON ATMOSPHERIC OXIDIZING CAPACITY

Pyrogenic emissions of ozone precursors are abundant (Andreae, 1991), and ozone formation from biomass fires has been the subject of much study (Granier et al., 1996; Lelieveld et al., 1997). The steps in ozone formation are the same as smog reactions in urban environments, although non-gas-phase chemistry may also play a role because particulate emissions from fires are substantial. The release of reactive hydrocarbons (CH_4, but more importantly, nonmethane hydrocarbons), carbon monoxide and NO (nitric oxide) produces a mixture that enhances ozone formation. Table 2 shows the sequence of reactions with NO, CO, and nonmethane hydrocarbons (designated as RH).

TABLE 1 Campaigns with Significant Biomass Burning Observations

Date	Name (Acronym)	Location	Reference
1— 2—Jul–Aug., 1985 April–May, 1987 3—July–August, 1988, 1990	Atmospheric Boundary Layer Experiments (ABLE), 1, 2, 3	1—Tropical Atlantic Ocean 2—Brazilian Rain Forest 3—Alaskan northern wetlands	*JGR*[a] 93: (D2) Feb. 20 1988 *JGR* 95: (D10) Sep. 20 1990 *JGR* 97: (D15) Oct. 30 1992 *JGR* 99: (D1) Jan. 20 1994
I—1987 II—1991	TROPOZ	Europe to America to South Africa and return	I, Quad. Ozone—1988[b]
1988	DECAFE	Equatorial Africa	*JGR* 97: (D6), 6187–6193, 1992
1991	FOS		*J. Atmos. Chem.*, 22 (1), 1995
Aug.–Oct. 1992	TRAnsport and Chemistry near the Equatorial-Atlantic (TRACE-A)	South tropical Atlantic Ocean	*JGR* 101: (D19), 23515–24330, 1996
1992	Southern Africa Fire-Atmosphere Research Initiative (SAFARI-92)	Southern Africa	*JGR* 101: (D19) 23505–24330, 1996
May, 1994	Southern African Atmospheric Research Initiative (SA'ARI-94)	Southern Africa	*S. Afr. J. Sci.* 91: (7), 360–362, July 1995
May–June, 1996 Jan./Feb., 1997	EXPeriment for REgional Sources and Sinks of Oxidants (EXPRESSO)	Central African Republic and Republic of Congo	*JGR* 104: (D23) 30625–30657, Dec. 20 1999
Sep./Oct., 1997	SAFARI-97 Field Campaign in Kenya	Kenya	*JGR* 102: (D15) 18879–18888, Aug. 20 1997
Feb./March, 1998 October, 1999	Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA) CLAIRE	Amazon, Brazil; Surinam	*Ann. Geophys-Atm.-Hydr.* 17: (8) 1095–1110, Aug. 1999

Note. Those since 1990 were ICAC/BIBEX sponsored.

[a]JGR = *Journal of Geophysical Research.*

[b]Quad. Ozone—1988 = *Ozone in the Atmosphere*, R. D. Bojkov and P. Fabian, A. Deepak, Eds., Pub., Hampton, VA, 1989.

TABLE 2 Photochemical Reactions Linking Methane, NMHC, CO, and NO with O_3, OH

OH Forms from Ozone Photolysis

$O_3 + h\nu \rightarrow O(^1D) + O_2 \qquad O(^1D) + H_2O \rightarrow OH + OH$

Methane Oxidation

$OH + CH_4(+O_2) \rightarrow CH_3O_2 + H_2O$

$CH_3O_2 + NO \rightarrow NO_2 + CH_3O$

[Form formaldehyde: $CH_3O(+O_2) \rightarrow HCHO$]

$HCHO + h\nu \rightarrow H_2 + CO \leftarrow$ formation of CO

$HCHO + h\nu(+O_3) \rightarrow HO_2 + CHO$

NMHC Oxidation

$OH + NMHC(+O_2) \rightarrow RO_2 + H_2O$

$RO_2 + NO \rightarrow NO_2 + RO$

CO Oxidation by OH Produces HO_2

$OH + CO(+O_2) \rightarrow CO_2 + HO_2$

Conversion of NO to NO_2 by HO_2, RO_2, CH_3O_2

$CH_3O_2 + NO \rightarrow NO_2 + CH_3O$

$RO_2 + NO \rightarrow NO_2 + RO$

$HO_2 + NO \rightarrow NO_2 + OH$

Formation of O_3

$NO_2 + h\nu \rightarrow O + NO$

$O + O_2(+M) \rightarrow O_3 + M$

3 RESULTS OF TROPICAL FIELD CAMPAIGNS

Trace Gas Signatures and Ozone Photochemistry

In Brazil and Africa, experiments directed toward biomass burning (e.g., DECAFE, TRACE-A, SCAR-B, TROPOZ I and II, EXPRESSO) and biogenic emissions (ABLE 2, ABLE 3, LBA) have shown that both pyrogenic and biogenic emissions can lead to substantial ozone formation. Biogenic sources appear to be most important in the boundary layer, below canopy level, where soil NO emissions lead to ozone formation. This was seen during ABLE 2A (Jacob and Wofsy, 1988), where 1 ppbv NO built up near the surface producing > 15 ppbv O_3/day. In addition to NO, isoprene emissions were essential to ozone formation. During the SAFARI-92 experiment (September–October 1992), elevated NO levels over the savanna following precipitation (Harris et al., 1996; Zepp et al., 1996) signified biogenic emissions. Aircraft sampling showed that these higher NO signals extended well into the mixed layer and that they lasted 1 to 3 days. This source of NO could contribute to ozone formation, providing a significant fraction in tropical regions during the dry (burning) to wet (nonburning) transition (Swap et al., 1996).

Despite contributions from biogenic sources, persistently high ozone levels throughout the free tropical troposphere during the dry season usually originate from biomass burning and occasionally from urban areas. Figure 2 shows typical ozone and ozone precursor profiles in a region affected by biomass burning, during the October 6, 1992 TRACE-A flight over Zambia. Figure 3 summarizes mean

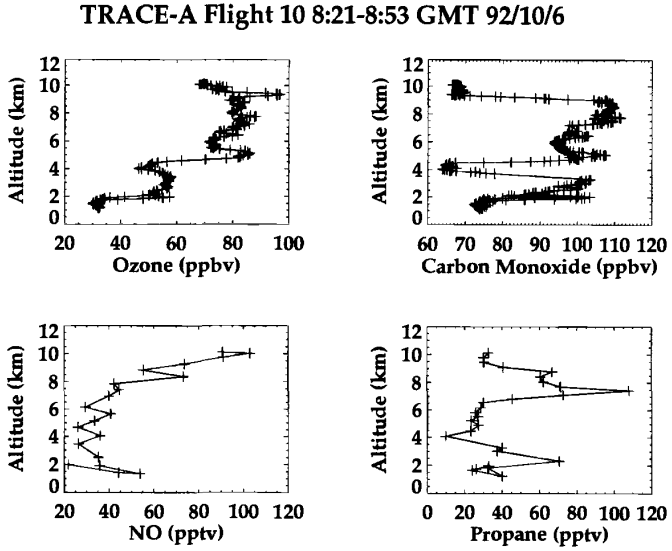


Figure 2 Profiles of ozone and ozone precursors CO, NO, and propane from sampling on board the NASA/DC-8 during the TRACE-A field experiment (on October 6, 1992, flight from Johannesburg to Zambia. Data available from: *http://www-gte.larc.nasa.gov.*)

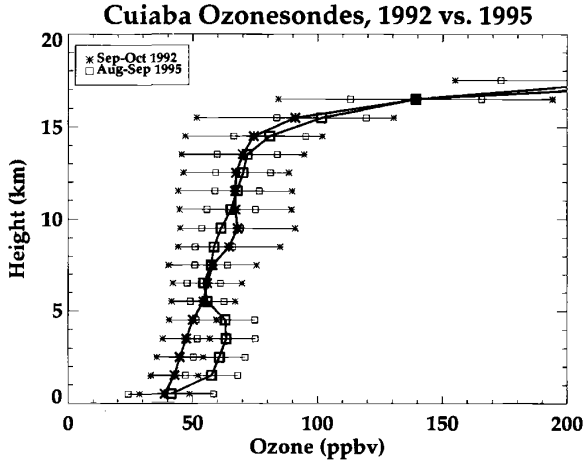


Figure 3 Ozone profiles from ozonesondes at Cuiabá, Brazil (16°S, 56°W) during biomass burning field experiments of 1992 (TRACE-A), which was a relatively low fire activity year and 1995 (SCAR-B), with greater burning activity. Mean 1-km profiles (to ±1-sigma) shown. Data from V. W. J. H. Kirchhoff.

profiles from ozonesondes launched at Cuiabá (Brazil, 16°S, 56°W) during two campaigns: TRACE-A and SCAR-B. Biomass burning (Artaxo et al., 1998) was much lower in 1992 (TRACE-A) than in 1995 (SCAR-B), and the boundary layer was more stable during the latter period; hence ozone levels in the lower troposphere were greater during SCAR-B.

Studies of ozone photochemical formation during sampling periods on TRACE-A have been made with photochemical steady-state ("point") models (Jacob et al., 1996; Thompson et al., 1996; Zenker et al., 1996; Mauzerall et al., 1998). The mixed layer, near the surface, usually has net ozone formation. For example, in the 4 km nearest the surface, relatively fresh emissions sampled over Brazil (September 27, 1992) and Zambia (October 6, 1992) during TRACE-A produced 10 to 15 ppbv O_3/day. During TROPOZ I, near the Ivory Coast in December 1987, air parcels with emissions less than 2 days old formed ozone at a 15 to 35 ppbv O_3/day rate (Jonquières et al., 1998). In PEM–Tropics A, near biomass burning in southeast Asia, near-surface ozone formation averaged >6 ppbv ozone/day (Schultz et al., 1999). Ozone formation increased with altitude because NO was supplied by peroxyacetylnitrate (PAN) transported into the region (Schultz et al., 1999). In contrast, during TRACE-A, above the mixed layer, ozone formation was in balance between production and loss or net negative during TRACE-A (Jacob et al., 1996; Thompson et al., 1996) because NO was depleted.

In the upper troposphere, photochemical formation often proceeds at modest rates (1 to 3 ppbv/day) and the ozone photochemical lifetime is 2 weeks to 2 months (Thompson et al. 1996; Jacob et al., 1996; Schultz et al., 1999). Usually, ozone formation is slightly positive because NO concentrations are sufficiently high. The requisite NO concentrations (>50 pptv) may come from lightning enhancement of NO, recycling of reactive nitrogen, or convective injection of NO_x. Venting of the boundary layer in convective cells produced by the intense heat of biomass fires is another mechanism whereby ozone precursors are injected into the free troposphere (Chatfield et al., 1996; 1998). This was seen during African aircraft sampling in TROPOZ (Jonquières et al., 1998) and TRACE-A (Thompson et al., 1996; Mauzerall et al., 1998).

Analyses of ozone formation show its evolution during transit away from the continents of biomass burning (Chatfield et al., 1996; Thompson et al., 1996; Jonquières and Marenco, 1998; Jonquières et al., 1998; Mauzerall et al., 1998; Schultz et al., 1999). Mauzerall et al. (1998) classified the age of air in terms of reactive nitrogen species (NO, HNO_3, PAN) and CO content, using ratios of tracers like CO_2, C_2H_2, C_2H_6, and CH_3COCH_3. In parcels sampled during TRACE-A, the limiting ozone-forming reactant was NO, which tended to be used up within a day or so. Older air parcels are refreshed with NO as PAN decomposes thermally, releasing NO_2, which is rapidly photolyzed to NO. Thus, downwind, ozone formed from NO that was supplied by PAN.

Schultz et al. (1999) examined photochemical characteristics of African plumes, observed thousands of kilometers from their sources, over the Pacific during the September PEM–Tropics A expedition. Figure 4 shows CO over the tropical Pacific (Blake et al., 1999), with many elevated CO segments due to pyrogenic sources from several continents (Olson et al., 1999). Schultz et al. (1999) also find the PAN mechanism for NO to be dominant, as in TRACE-A, but advection of ozone, not

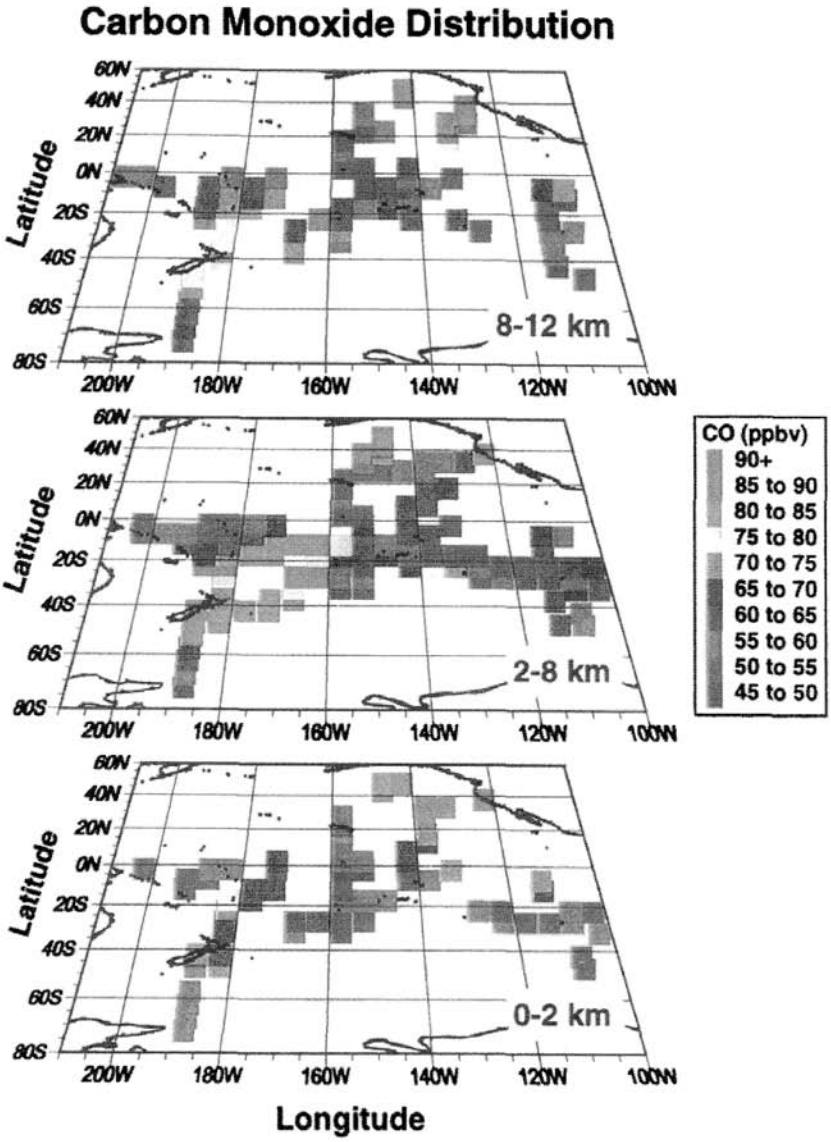


Figure 4 (see color insert) CO over tropical Pacific during September 1996 PEM–Tropics A sampling (from Blake et al., 1999). Measurements by G. W. Sachse with a lidar-based instrument. Analysis of possible fire sources is described by Olson et al. (1999). See ftp site for color image.

local photochemistry, is still a major tropical Pacific ozone source. A plume of ozone with African origins observed over the western Pacific appears in Figure 5.

Biomass burning is not the only large nonurban NO source that contributes to tropical ozone formation. Lightning is also a significant source. TRACE-A observa-

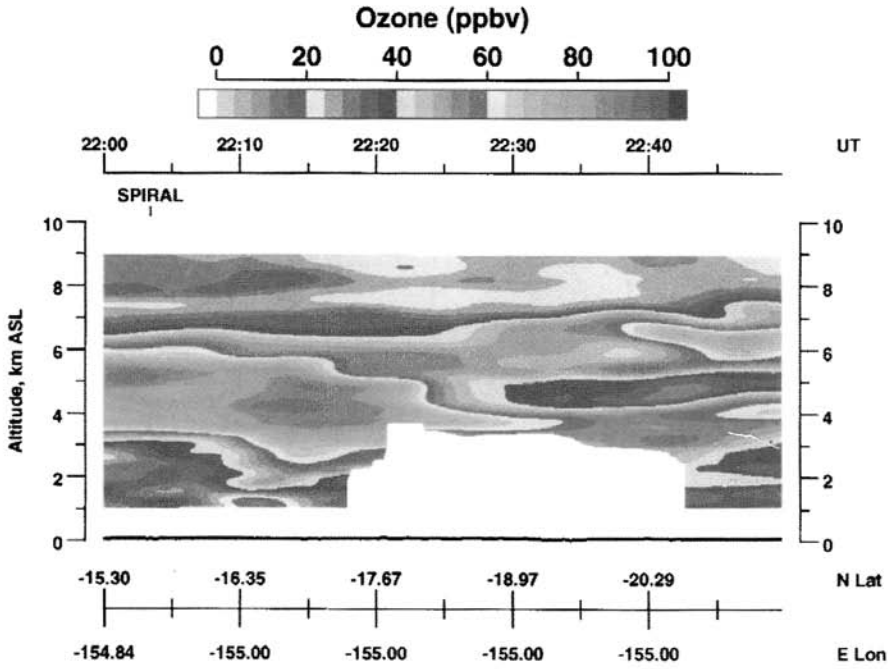


Figure 5 (see color insert) Ozone plume over the Pacific seen during the PEM–Tropics A aircraft mission in Sept.–Oct. 1996. (from Fenn et al., 1999). See ftp site for color image.

tions throughout the south tropical Atlantic, for example, showed elevated, relatively fresh NO in the upper tropospheric that did not always track other tracers of biomass burning (Smyth et al., 1996). A TRACE-A flight (September 27, 1992) over Brazil in which deep convection transported relatively fresh biomass burning emissions to the upper troposphere (Pickering et al., 1996) was also punctuated by lightning-produced NO. From comparison of NO enhancements to other biomass burning emittants (CO, hydrocarbons), it appeared that 35 to 40% of the NO from the September 27, 1992 flight on TRACE-A was due to lightning.

Transport of Trace Gas Emissions and Ozone from Biomass Burning

Aircraft, ground-based and sounder sampling, in combination with trajectory and regional dynamical modeling, elucidates the roles of convection and long-range transport in determining the distribution of smoke aerosol, tropical ozone, and ozone precursor distributions. Over large biomass burning regions of north equatorial Africa, the Harmattan winds cause large-scale transport of biomass burning products from the continent in a southwesterly flow to the Atlantic and toward South America (Jonquières and Marenco, 1998; Fig. 6*a*). From southern African savanna burning, convection on a large regional scale drives a Southern Hemisphere "Great Plume" (Chatfield et al., 1996, 1998) toward the tropical Atlantic and Indian

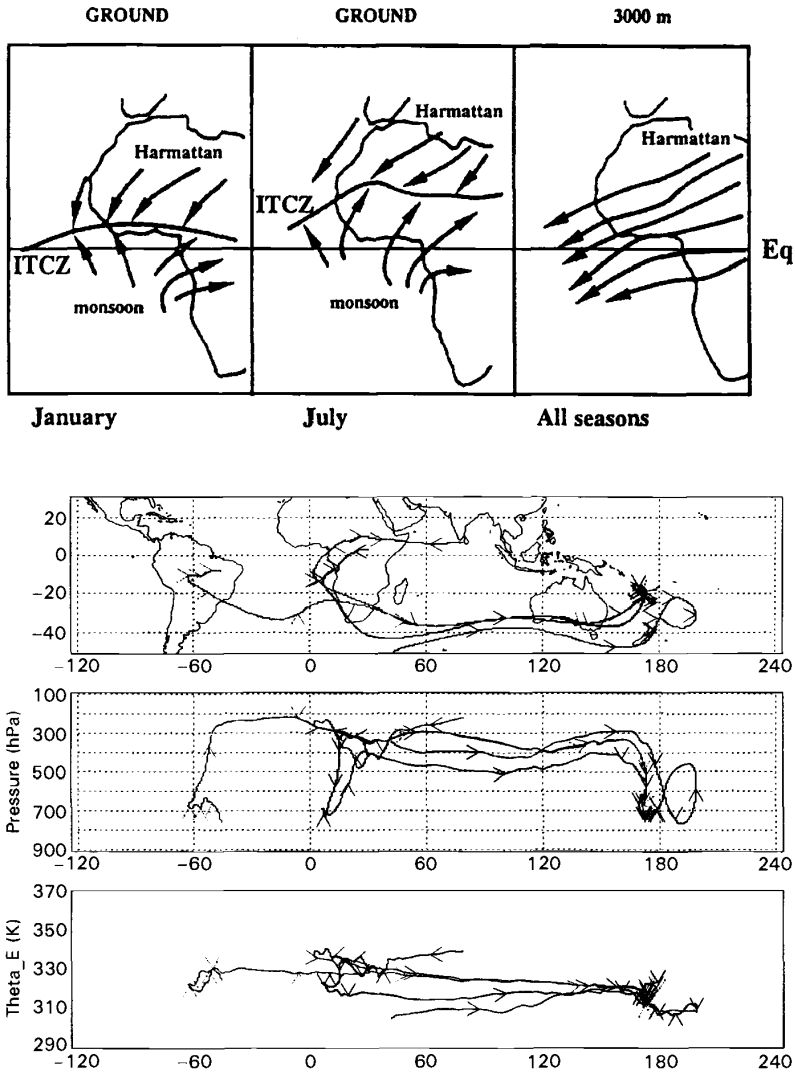


Figure 6 (*a*) Transport patterns from north equatorial Africa (Jonquières et al., 1998); (*b*) Aspects of the "Southern Global Plume" from Guo and Chatfield (1998). Back trajectories are initiated at several points along the NASA DC-8 flight of September 23–24, 1996 (PEM–Tropics A). Top panel shows traces back to origin areas in South America and Africa. Middle panel shows pressures of trajectories. Directional arrows are spaced every 2 days. Lowest panel shows potential temperature of trajectory, which is nearly conserved, demonstrating that trajectory does not cross into other air masses. Model is based on MM5 mesoscale model. See ftp site for color image.

Oceans. Figure 6*b* shows that fires from South America can also affect the Indian and Pacific Oceans.

Ozone from both southern Africa and South America appears seasonally over the south Atlantic in tropospheric ozone satellite retrievals (Chapter 5). South of 15°S, the predominant exit for biomass burning emissions and ozone, which can accumulate in stable layers (Garstang et al., 1996; Garstang and Tyson, 1997; Tyson et al., 1997) is toward the Indian Ocean. In both Atlantic and Indian Ocean exit routes from southern Africa, emissions from fires, vented by shallow or deep convection, inject most of the ozone precursors into the 4 to 8 km layer. These are readily detected in ozone profiles from balloon-borne sondes released over Réunion Island (21°S, 55°E; Baldy et al., 1996; Randriambelo et al., 1999). Examples of fire-affected layers at Réunion appear in Figure 7. In Guo and Chatfield (1998), tracers in the 5th version PennState/NCAR Mesoscale Model (MM5) simulate the flow of CO from industrial, biogenic, and biomass burning sources over southern Africa to the western Pacific Ocean. CO mixing ratios computed by the model agree with observations during PEM–Tropics A (Hoell et al., 1999).

The route of biomass burning emissions from Brazil has been studied on ABLE 2A, TRACE-A and SCAR-B. Figure 8, which is based on a composite of forward trajectories during the SCAR-B experiment (Longo et al., 1999), shows air parcels

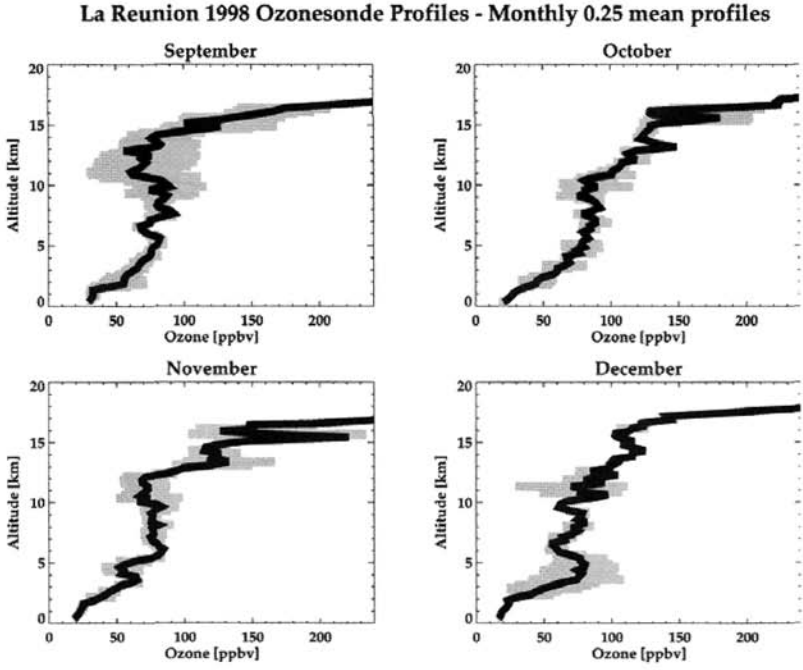


Figure 7 Ozone soundings over Réunion Island (21°S, 55°E) in the Indian Ocean, with layers of high ozone due to transport from African burning. Mean monthly profiles with l-sigma shading.

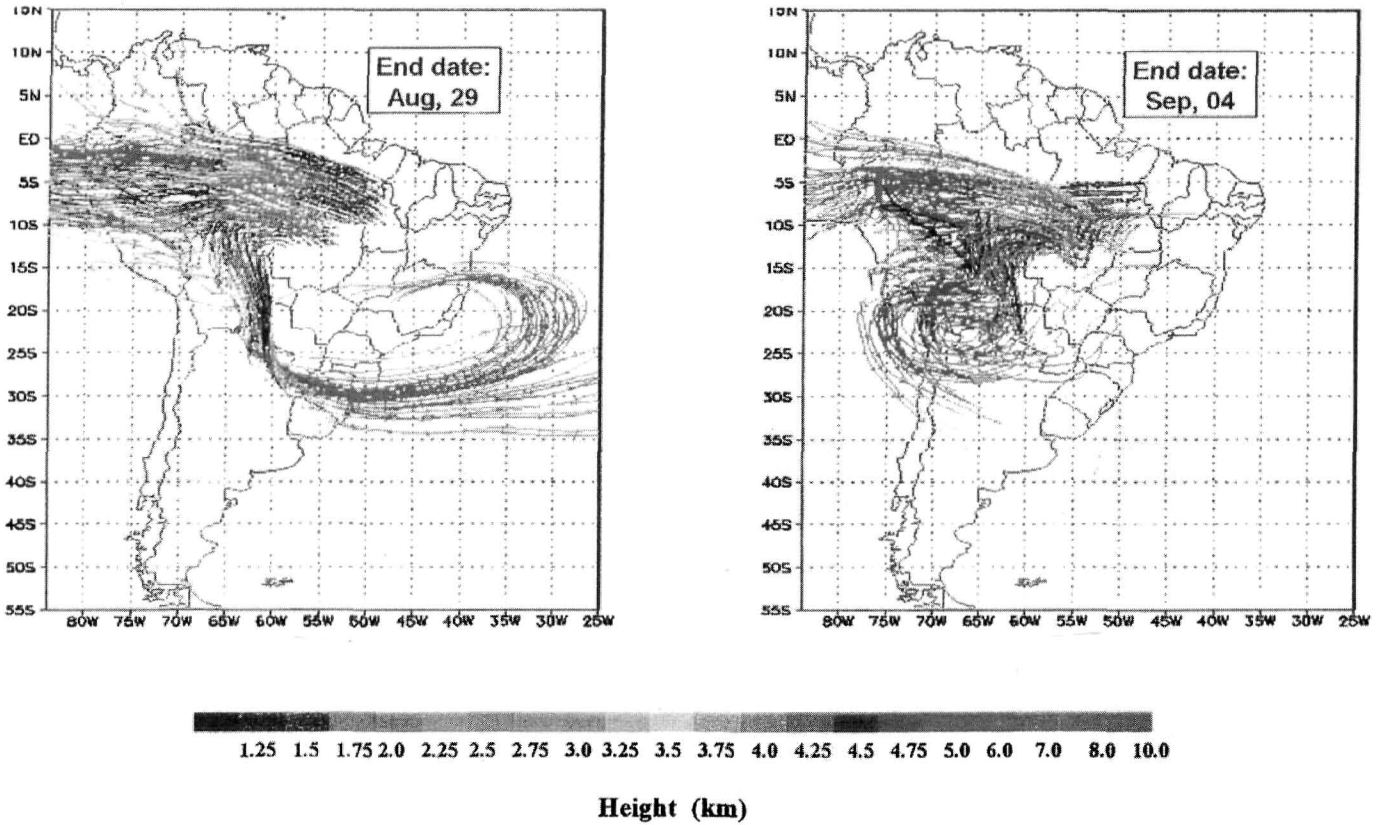


Figure 8 (see color insert) Composite of forward trajectories from Cuiabá during the 1995 SCAR-B field experiment. A Brazilian version of the Colorado State mesoscale RAMS model was used to provide winds for the University of São Paulo kinematic trajectory model (from Longo et al., 1999). See ftp site for color image.

from active burning regions sending ozone and ozone precursors over mountains toward the eastern Pacific Ocean. Unfortunately, there was no satellite remote sensor available in August 1995 to detect flows over the eastern Pacific during SCAR-B. However, satellite data from 1979 to 1992 (Kim and Newchurch, 1996; Thompson and Hudson, 1999) show a seasonal drift of ozone into this region. During TRACE-A, the predominant post-convective flow from Brazilian biomass burning areas at the onset of the wet season was in the westerlies toward the Atlantic. It was estimated that upper tropospheric ozone was largely supplied from the South American continent (Thompson et al., 1997), with additional ozone resulting from lightning-produced NO.

Ozonesondes over Africa and South America, near or downwind from sources, as well as ozonesondes at more remote locations—Réunion (21°S, 55°E), Ascension (8°S, 14°W), American Samoa (14°S, 170°W)—show impacts of biomass burning ozone (Cros et al., 1992; Fishman et al., 1992; Baldy et al., 1996; Oltmans et al., 1998). Examples of ozone profiles at Pretoria (25°S, 28°E) and Etosha Park (19°S, 15°E) during SAFARI-92 and TRACE-A appear in Figures 9*a* and 9*b*. Neither of these sites is in a burning region, but clusters of back trajectories initiated at the peaks with arrows show that they may be several days' transport time from African burning or a week from South American savanna burning. Back trajectories from the Etosha Park ozonesonde profile of October 11, 1992, indicated significant exposure to fires within 2 days (Fig. 9*c*). For the Pretoria ozonesonde sample on October 11, 1992, launched within 30 km of Johannesburg, the number of fires encountered in a 5-day back trajectory is less and travel time is greater than air parcel origins on October 11, 1992, at Etosha Park.

Airborne sampling and sounding profiles show that layers of enriched or depleted ozone are remarkably stable (Garstang et al., 1996; Garstang and Tyson, 1997; Newell et al., 1999). Using the SAFARI-92/TRACE-A soundings over Irene, Garstang et al. (1996) found very little vertical mixing and estimated that some of the stable layers observed had lifetimes greater than 50 days.

4 REMOTE SENSING

Remote sensing is an invaluable tool for looking more closely at biomass burning effects in the atmosphere. In terms of trace gases, ozone and CO instrumentation flown on aircraft and in space has seen the imprint of biomass fires on a widespread basis. Remote sensing of carbonaceous absorbing aerosols (soot, smoke) is made from airborne platforms and from a number of space-borne instruments. In addition, imagers are able to detect fires (Cahoon et al., 1992) and fire burn scars on the earth (Justice et al., 1996). Instrumentation is summarized in Table 3 and applications are described below.

Carbon Monoxide

The Shuttle-borne Measurement of Air Pollution by Space (MAPS; Reichle et al., 1990; *Journal of Geophysical Research*, Aug. 20, 1998) instrument is a gas correla-

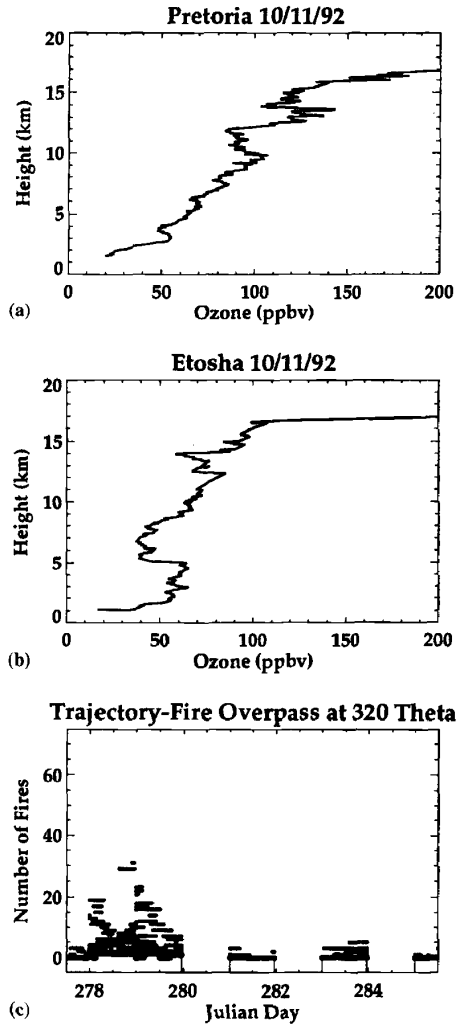


Figure 9 (*a*) Pretoria ozone sounding for October 11, 1992. (*b*) Etosha sounding for October 11, 1992. (*c*) Fires passed over by air parcels in a cluster of back trajectories initiated at ~5 km ($\theta = 320$ K) at the Etosha location on October 11, 1992. This suggests fire emissions contribution to ozone profile in (*b*). Satellite fire counts from Justice et al. (1996); gaps refer to days with missing fire data.

TABLE 3 Remote Sensing Instrumentation for Detection of Smoke, Fires, and Trace Gas Emissions

MAS: airborne surface imager
AVHRR, GOES: smoke detection from fires
AVHRR: surface imaging for active fires and burn scars
Cloud and smoke lidar: NASA/ER-2 instrument
DMSP: active fires
GOME: ozone, NO_2, HCHO, BrO
MAPS, MOPITT: CO
TOMS: ozone, smoke, and dust aerosol (also SO_2, sulfate aerosols)

tion radiometer that senses CO by differencing two cells. For the region of the atmosphere of greatest sensitivity, between 5 and 10 km, MAPS gives an accurate measurement of carbon monoxide. Operating on the Space Shuttle in 1981, twice in 1984 and in 1994, with data covering 55°N to 55°S, MAPS observed CO from urban pollution as well as from biomass burning. Biomass burning signatures in the tropics

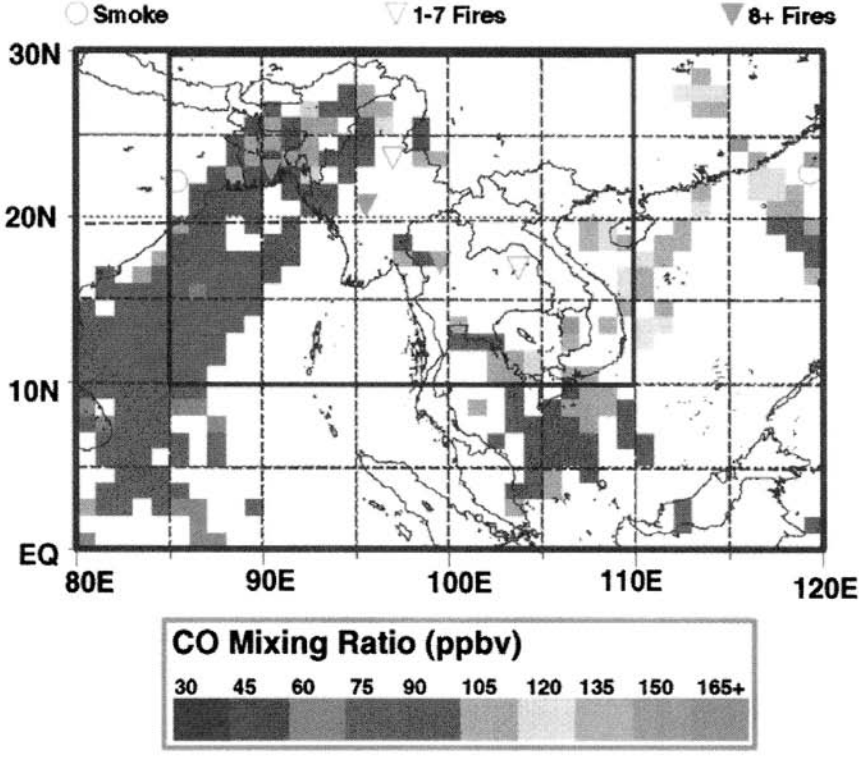


Figure 10 (see color insert) (*a*) MAPS CO, April 1994 (from Christopher et al., 1998). See ftp site for color image.

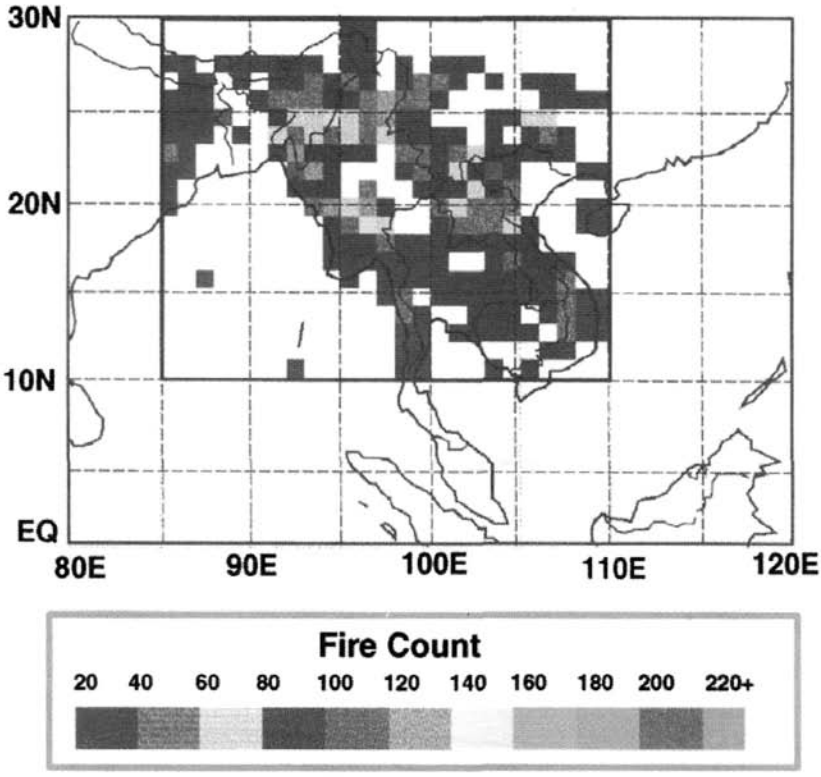


Figure 10 (see color insert) (*b*) coincident fires during April 1994 Space Shuttle flight (from Christopher et al., 1998). See ftp site for color image.

are evident as elevated CO concentrations, usually >60 ppbv. This was confirmed in an airborne campaign of validation measurements conducted during the 1994 MAPS operations. Because MAPS detects midtropospheric CO, it essentially detects areas of burning and convective transport in which CO from the boundary layer is transported to midtroposphere. Urban CO that escapes the boundary layer can also be detected. Figure 10 shows MAPS CO and remotely sensed fires contributing to the CO over southern Asia (from Christopher et al., 1998).

MOPITT, the new CO and methane sensor aboard the *Terra* spacecraft, was launched into orbit in December 1999. As of this writing, MOPITT observations had not yet begun.

Tropospheric Ozone

The application of ozone remote sensing to the troposphere, in a series of studies by Fishman and co-workers (Fishman et al., 1986, 1990; Fishman and Brackett, 1997),

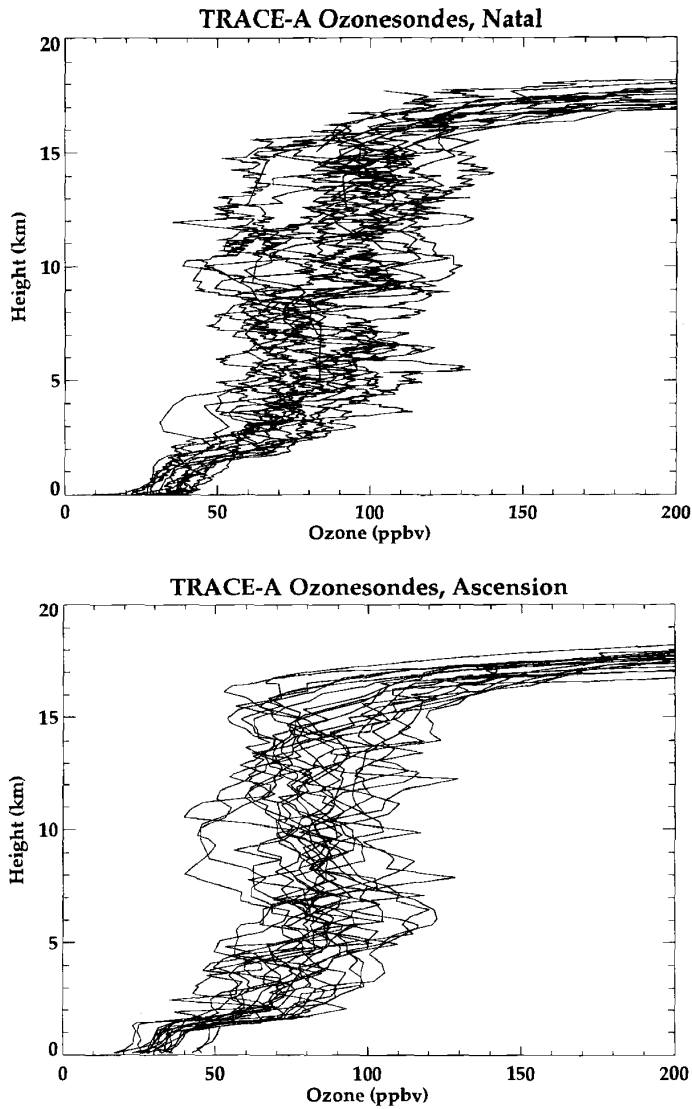


Figure 11 Ozonesonde profiles during the TRACE-A field experiment (September 9 to October 22, 1992) over (*a*) Natal, Brazil (6°S, 35°W) and (*b*) Ascension Island (8°S, 15°W).

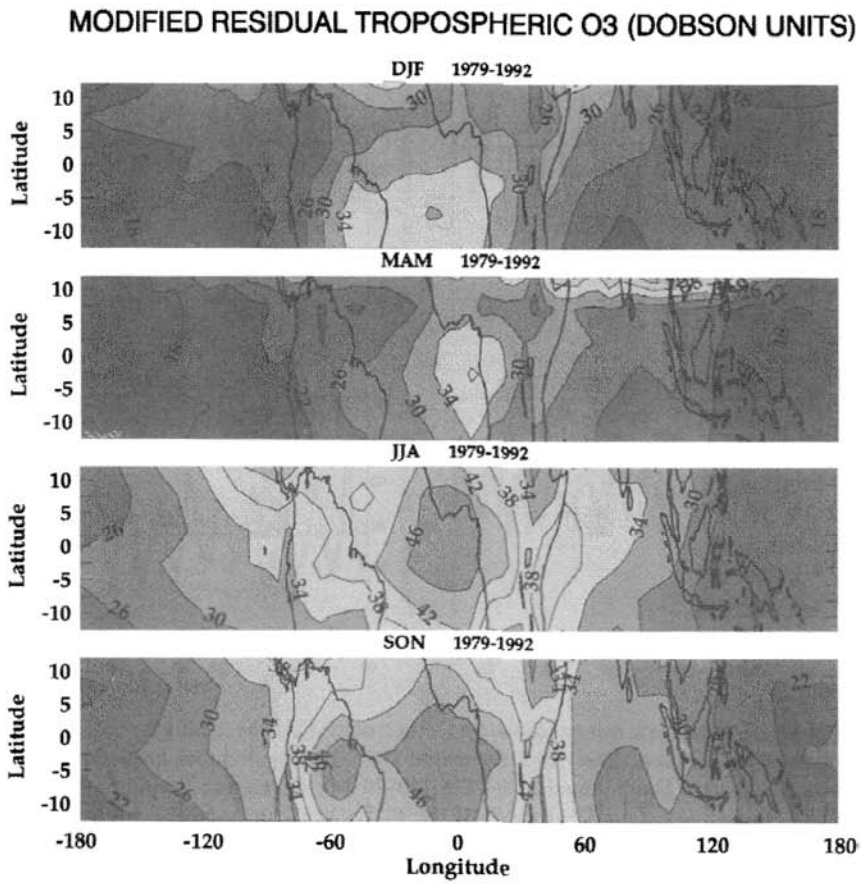


Figure 12 (see color insert) Wave-one pattern in tropospheric ozone apparent in TOMS satellite data, averaged from 2 maps/month during the 1979–1992 *Nimbus 7* observing period. Wave appears to be present throughout year. Scale is DU (Dobson units). Cf. Figure A1 in Thompson and Hudson (1999). See ftp site for color image.

gave the first insight into the extent of biomass burning effects on tropospheric ozone. The entire south Atlantic basin shows a tropospheric ozone maximum in the latter part of the Southern Hemisphere biomass burning season. Because the TOMS satellite instrument measures column ozone, and has limited sensing capacity below 500 mbar, the vertical characteristics of enhanced ozone seen from space had to be confirmed by ozonesondes (Fishman et al., 1992). Layering of ozone from the boundary layer to 15 km is evident in sondes from Natal (coastal Brazil at 6°S) and Ascension Island (8°S, 15°W; Fig. 11). These profiles were taken during the 1992 SAFARI/TRACE-A experiments.

The intensity of the ozone maximum feature varies from year-to-year, and the chemical consequences of biomass burning appear to overlie a persistent wave-one

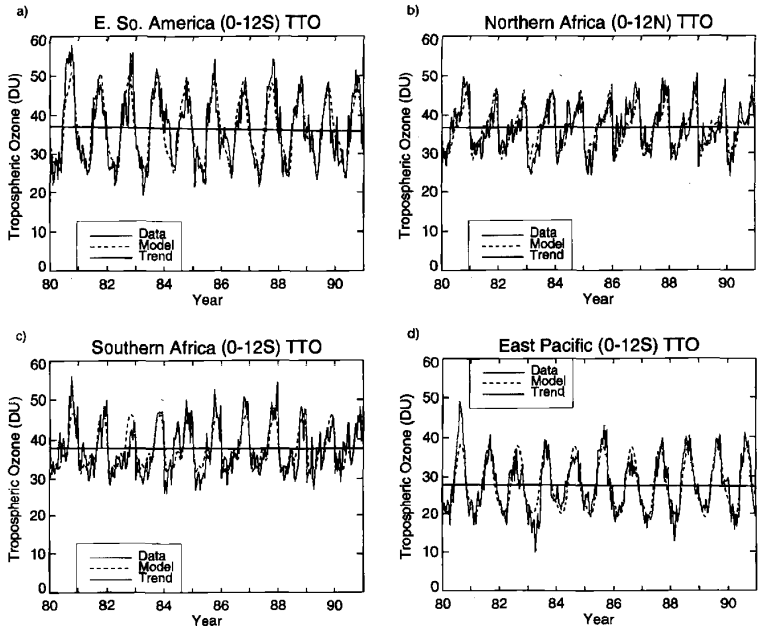


Figure 13 Tropospheric column ozone (in Dobson units) from the modified-residual method (Thompson and Hudson, 1999) over the period 1980–1990 within four tropical regions as follows: (*a*) eastern South America (0–12°S, 40–70°W); (*b*) eastern Pacific (0–12°S, 80–110°E); (*c*) northern equatorial Africa (0–12°N, 20°W–20°E); (*d*) southern equatorial Africa (0–12°S, 0–30°E). Data available at *metosrv2.umd.edu/~tropo/*

pattern that maintains nonpollution Atlantic tropical tropospheric ozone at an always greater column depth than nonpollution Pacific ozone (Hudson and Thompson, 1998; Thompson and Hudson, 1999; Ziemke et al., 1996). An example of wave-one patterns in tropospheric ozone, taken from tropical tropospheric ozone maps, appears in Figure 12. During the Southern Hemisphere dry season, tracers of savanna fires and photochemical analysis show that elevated ozone over the south Atlantic basin is dominated by biomass burning sources; see references for TROPOZ, DECAFE, SAFARI, TRACE-A, and SCAR-B (Table 1). This ozone amounts to 20 to 30 DU (Dobson unit) more over the Atlantic region than over the Pacific at the same season and 20 to 30 DU more than is over the Atlantic during its seasonal minimum (March–May). These features are apparent in a time series of TOMS-based tropospheric ozone data (Thompson and Hudson, 1999; Ziemke et al., 1998). Mean annual tropospheric ozone column in eastern South America and Africa (deseasonalized value given by straight lines in Figs. 13*a* to 13*c*) is 38 DU compared to 28 DU over the eastern Pacific (Fig. 13*d*).

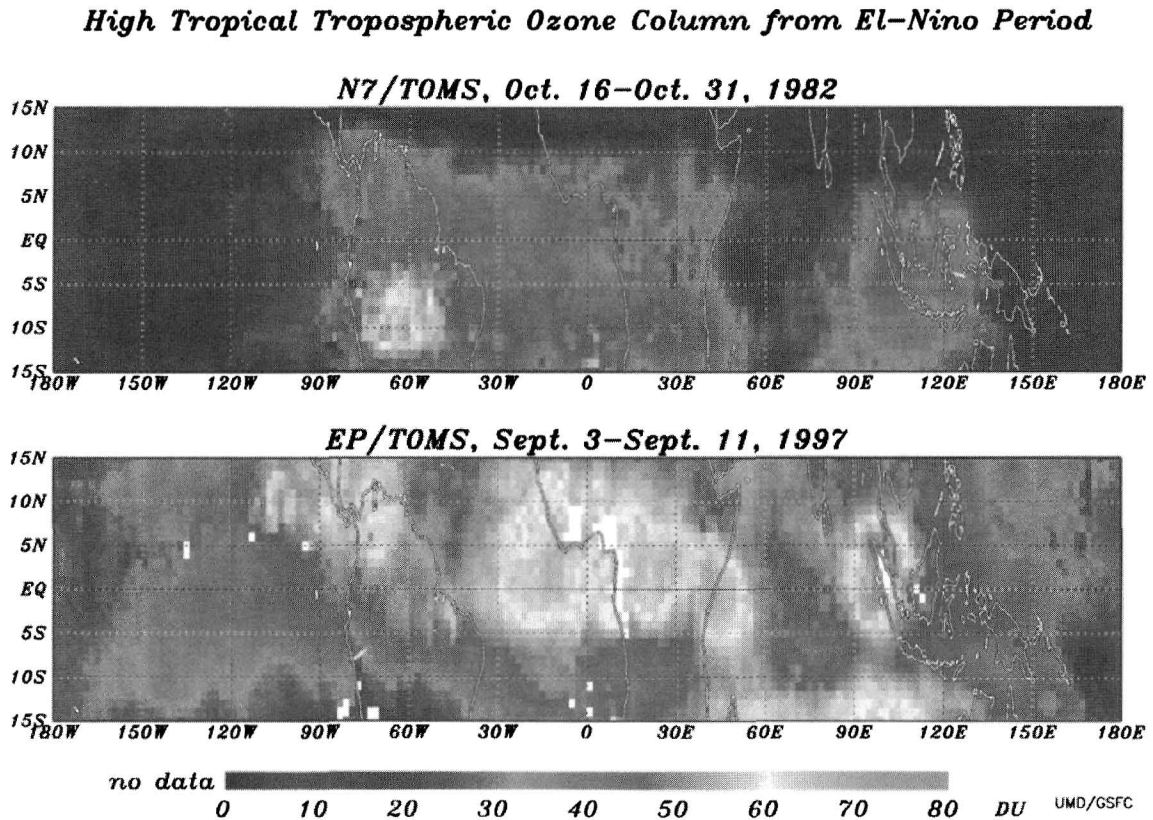


Figure 14 (see color insert) Tropospheric column ozone (in DU, from modified-residual method; Thompson and Hudson, 1999) during El Niño-Southern Oscillation (ENSO) of late 1982 (upper panel) as seen in tropical tropospheric ozone map and for September 1997 (lower panel). See ftp site for color image.

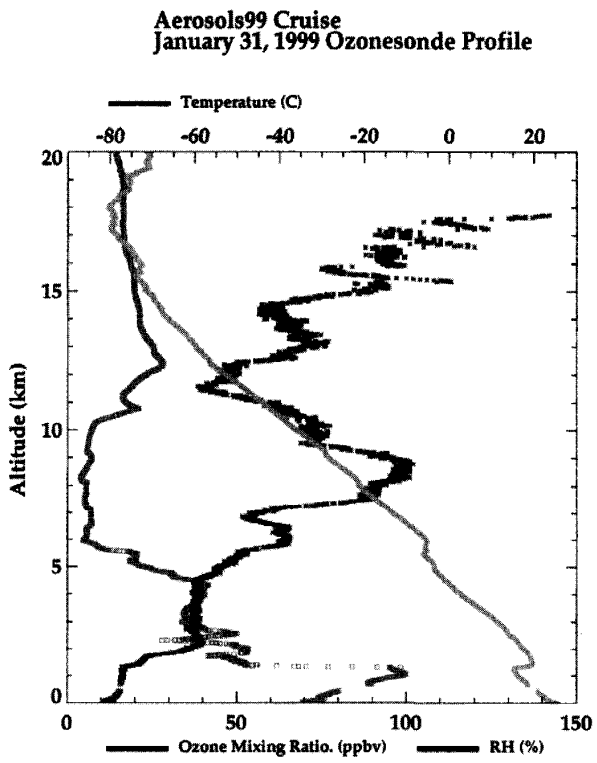


Figure 15 (see color insert) (*a*) Profiles of ozone, temperature and water vapor (as percent relative humidity) from 0 to 20 km on January 31, 1999 during Aerosols99 cruise of R/V *Ronald H. Brown*. Anti-correlation of high ozone between 7 and 10 km suggestive of aged stratospheric air. (*b*) Comparison of integrated tropospheric column ozone from sondes launched along Atlantic transect of R/V *Ronald H. Brown* (Thompson et al., 2000) in January–February 1999 and from sondes launched along January–February 1993 Atlantic transect of R/V *Polarstern* (Weller et al., 1996). See ftp site for color image.

The TOMS-based maps are used to characterize interannual variability and seasonality of tropical ozone. In Figure 13, two features stand out. One is that over the 11-year period illustrated, there is no significant trend in tropospheric ozone (Thompson and Hudson, 1999; Chandra et al., 1999), despite an apparent increase in smoke aerosols in some of these areas (Hsu et al., 1999). The second noteworthy feature is the presence of extremes in tropospheric ozone during the strong El Niño episode of late 1982 and early 1983. For example, over South America there was elevated ozone due to higher-than-usual biomass burning activity (Fig. 14, upper panel), whereas over the eastern Pacific, with enhanced convective activity, upward transport of ozone diluted (and reduced) column ozone. Very high ozone and biomass burning aerosol signals were observed by TOMS over the

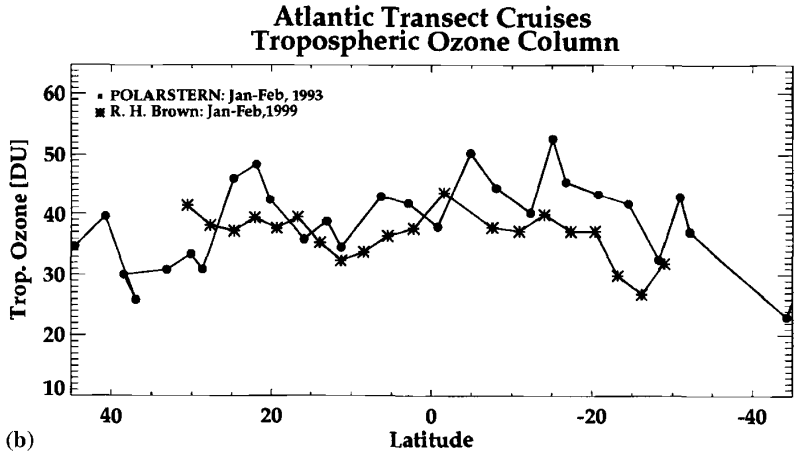


Figure 15b

Indonesian region (Fig. 14, lower panel) during the 1997–1998 El Niño event. Ozonesondes in Java, downwind of the most intensely burning regions of Indonesia (Liew et al., 1999) also registered high tropospheric column ozone (Fujiwara et al., 1999).

Krishnamurti et al. (1996) showed that ozone accumulates over the south Atlantic due to dynamical forces that tends to produce an Atlantic–Pacific ozone gradient (more ozone over the Atlantic) irrespective of chemical sources. Evidence of dynamical effects are easiest to isolate in ozonesondes recorded over the tropical Atlantic during the Southern Hemisphere wet season. For example, Atlantic oceanographic transects with ozonesonde launches (Smit et al., 1989; Weller et al., 1996; Thompson et al., 2000) have shown free tropospheric ozone in the Southern Hemisphere dominated by high ozone layers that may originate from cross-hemispheric transport (Jonquières and Marenco, 1998) or aged air parcels from the stratosphere (the latter inferred from very low water vapor, Fig. 15*a*). The result is a paradox with respect to biomass burning in that there is more tropospheric ozone in the Southern Hemisphere wet season than there is north of the Intertropical Convergence Zone (Fig. 15*b*), where there is active biomass burning over northern equatorial Africa.

ACRONYMS

ABLE	Amazon Boundary Layer Experiment (A = 1985; B = 1987)
AVHRR	Advanced Very High Resolution Radiometer
BIBEX	Biomass Burning Experiment
CLAIRE	Coordinated LBA (Large Basin Amazonia) Atmospheric Experiment

DECAFE	Dynamique Et Chimie Atmosphérique en Forêt Equatoriale [1988; FOS (Fires of Savannas)/DECAFE = 1991]
DMSP	Defense Mapping Satellite Project
EXPRESSO	Experiment for Regional Sources and Sinks of Oxidants (1996)
GOME	Global Ozone Monitoring Experiment (operating 1995–)
IGAC	International Global Atmospheric Chemistry Project
MAPS	Measurements of Air Pollution from Shuttle (1981, 1984, 1994)
MAS	MODIS (Moderate Resolution Imaging Spectrometer) Airborne Simulator
MOPITT	Measurements of Pollution in the Troposphere
PEM–Tropics A	Pacific Exploratory Mission (1996)
PEM–Tropics B	Pacific Exploratory Mission (1999)
SAFARI	Southern African Fire Atmospheric Research Initiative (1992)
SCAR-B	Smoke, Clouds and Radiation—Brazil (1995)
SEAFIRE	Southeast Asia Fire Experiment (1997)
TOMS	Total Ozone Mapping Spectrometer (*Nimbus 7*, 1978–1993; *Meteor*, 1991–1994; *ADEOS*, 1996–1997; *Earth-Probe*, 1996–)
TRACE-A	Transport and Atmsopheric Chemistry near the Equator—Atlantic (1992)
TROPOZ	Tropospheric Ozone Campaigns (I = 1987; II = 1991)

ACKNOWLEDGMENTS

Thanks to Robert Chatfield and Volker Kirchhoff for comments, discussion, and prepublication results. Graphical and manuscript assistance were provided by J. C. Witte, T. L. Kucsera (SSAI at NASA/Goddard), A. V. Cresce (University of Maryland), and J. R. Ziemke (SCA at NASA/Goddard).

REFERENCES

Andreae, M. O., Biomass burning: Its history, use and distribution and its impact on environmental quality and global climate, in J. S. Levine (Ed.), *Global Biomass Burning: Atmospheric, Climatic and Biospheric Implications*, MIT Press, Massachusetts, 1991, pp. 3–21.

Artaxo, P., E. T. Fernandes, J. V. Martins, M. A. Yamasoe, P. V. Hobbs, W. Maenhaut, K. M. Longo, and A. Castanho, Large-scale aerosol source apportionment in Amazonia, *J. Geophys. Res.*, *103*, 31837–31848, 1998.

Baldy S., G. Ancellet, M. Bessafi, A. Badr, and D. Lan Sun Luk, Field observations of tropospheric vertical distribution of tropical ozone at a remote marine site in the southern hemisphere, *J. Geophys. Res.*, *101*, 23835–23849, 1996.

Blake, N. J., D. R. Blake, O. W. Wingenter, B. C. Sive, L. M. McKenzie, J. P. Lopez, I. J. Simpson, H. E. Fuelberg, G. W. Sachse, B. E. Anderson, G. L. Gregory, M. A. Carroll, G. M. Albercook, and F. S. Rowland, Influence of southern hemispheric biomass burning on mid-tropospheric distributions of nonmethane hydrocarbons and selected halocarbons on the remote South Pacific, *J. Geophys. Res.*, *104*, 16213–16232, 1999.

Cahoon, Jr., D. R., B. J. Stocks, J. S. Levine, W. R. Cofer III, and K. P. O'Neill, Seasonal distribution of African savanna fires, *Nature*, *359*, 812–815, 1992.

Chandra, S., J. R. Ziemke, and R. W. Stewart, An 11-year solar cycle in tropospheric ozone from TOMS measurements, *Geophys. Res. Lett.*, *26*, 185–188, 1999.

Chatfield, R. B., J. A. Vastano, H. B. Singh, and G. W. Sachse, A general model of how fire emissions and chemistry produce African/Oceanic plumes (O_3, CO, PAN, smoke) seen in TRACE-A, *J. Geophys. Res.*, *101*, 24279–24306, 1996.

Chatfield, R. B., J. A. Vastano, L. Li, G. W. Sachse, and V. S. Connors, The Great African plume from biomass burning: Generalizations from a three-dimensional study of TRACE A carbon monoxide, *J. Geophys. Res.*, *103*, 28059–28077, 1998.

Christopher, S. A., C. Joyce, and R. M. Welsh, Satellite investigations of fire, smoke, and carbon monoxide during April 1994 MAPS mission: Case studies over tropical Asia, *J. Geophys. Res.*, *103*, 19327–19336, 1998.

Cros, B., D. Nganga, A. Minga, J. Fishman, and V. Brackett, Distribution of tropospheric ozone at Brazzaville, Congo, determined from ozonesonde measurements, *J. Geophys. Res.*, *97*, 12869–12875, 1992.

Crutzen, P. J., and J. G. Goldammer, *Fire in the Environment: The Ecological, Atmospheric, and Climatic Importance of Vegetation Fires: Report of the Dahlem Workshop*, Wiley, New York, 1993.

Fenn M. A., E. V. Browell, C. F. Butler, W. B. Grant, S. A. Kooi, M. B. Clayton, G. L. Gregory, R. E. Newell, Y. Zhu, J. E. Dibb, H. E. Fuelberg, B. E. Anderson, A. R. Bandy, D. R. Blake, J. D. Bradshaw, B. G. Heikes, G. W. Sachse, S. T. Sandholm, H. B. Singh, and R. W. Thornton, Ozone and aerosol distributions and air mass characteristics over the South Pacific during the burning season, *J. Geophys. Res.*, *104*, 16197–16212, 1999.

Fishman, J., V. G. Brackett, and K. Fakhruzzaman, Distribution of tropospheric ozone in the tropics from satellite and ozonesonde measurements, *J. Atmos. Terr. Phys.*, *54*, 589–597, 1992.

Fishman, J., and V. G. Brackett, The climatological distribution of tropospheric ozone derived from satellite measurements using version 7 Total Ozone Mapping Spectrometer and Stratospheric Aerosol and Gas Experiment data set, *J. Geophys. Res.*, *102*, 19275–19278, 1997.

Fishman, J., P. Minnis, and H. G. Reichle, Use of satellite data to study tropospheric ozone in the tropics, *J. Geophys. Res.*, *91*, 14451–14465, 1986.

Fishman, J., C. E. Watson, J. C. Larsen, and J. A. Logan, The distribution of tropospheric ozone determined from satellite data, *J. Geophys. Res.*, *95*, 3599–3617, 1990.

Fujiwara, M., K. Kita, S. Kawakami, T. Ogawa, N. Komala, S. Saraspriya, and A. Suripto, Tropospheric ozone enhancements during the Indonesian forest fire events in 1994 and in 1997 as revealed by ground–based operations, *Geophys. Res. Lett.*, *26*, 2147–2420, 1999.

Garstang, M. and P. D. Tyson, Atmospheric circulation, vertical structure and transport, in B. van Wilgen, M. Andreae, J. Goldammer, and J. Lindesay (Eds.), *Fire Southern African*

Savanna: Ecological and Atmospheric Perspectives, University of Witwatersrand Press, Johannesburg, 1997, Chapter 6.

Garstang M., P. D. Tyson, R. J. Swap, M. Edwards, P. Kållberg, and J. A. Lindesay, Horizontal and vertical transport of air over southern Africa, *J. Geophys. Res.*, *101*, 23721–23736, 1996.

Granier, C., W-M. Hao, G. Brasseur, and J-F. Müller, Land-use practices and biomass burning: Impact on the chemical composition of the atmosphere, in J. S. Levine (Ed.), *Biomass Burning and Global Change*, MIT Press, Massachusetts, 1996, pp. 140–148.

Guo, Z., and R. B. Chatfield, Meteorology of the Southern Global Plume: African and South American fires pollute the south Pacific, paper presented at the Sixth International Conference on Atmospheric Sciences and Application to Air Quality, Beijing, November 3–5, 1998.

Harris G. W., F. G. Wienhold, and T. Zenker, Airborne observations of strong biogenic NO_x emissions from the Namibian savanna at the end of the dry season, *J. Geophys. Res.*, *101*, 23707–23711, 1996.

Hoell J. M., D. D. Davis, D. J. Jacob, M. O. Rodgers, R. E. Newell, H. E. Fuelberg, R. J. McNeal, J. L. Raper, and R. J. Bendura, Pacific Exploratory Mission in the tropical Pacific: PEM-Tropics A, August–September 1996, *J. Geophys. Res.*, *104*, 5567–5583, 1999.

Hsu, C. N., J. R. Herman, O. Torres, B. N. Holben, D. Tanre, T. F. Eck, A. Smirnov, B. Chatenet, and F. Lavenu, Comparisons of the TOMS aerosol index with Sun-photometer aerosol optical thickness: Results and applications, *J. Geophys. Res.*, *104*, 6269–6279, 1999.

Hudson, R. D., and A. M. Thompson, Tropical tropospheric ozone (TTO) from TOMS by a modified-residual method, *J. Geophys. Res.*, *103*, 22129–22145, 1998.

Jacob, D. J., and S. C. Wofsy, Photochemistry of biogenic emissions over the Amazon forest, *J. Geophys. Res.*, *93*, 1477–1486, 1988.

Jacob, D. J., B. G. Heikes, S. M. Fan, J. A. Logan, D. L. Mauzerall, J. D. Bradshaw, H. B. Singh, G. L. Gregory, R. W. Talbot, D. R. Blake, and G. W. Sachse, Origin of ozone and NO_x in the tropical troposphere: A photochemical analysis of aircraft observations over the South Atlantic Basin, *J. Geophys. Res.*, *101*, 24235–24250, 1996.

Jonquières, I., and A. Marenco, Redistribution by deep convection and long-range transport of CO and CH_4 emissions from the Amazon basin, as observed by the airborne campaign TROPOZ II during the wet season, *J. Geophys. Res.*, *103*, 19075–19091, 1998.

Jonquières, I., A. Marenco, A. Maalej, and F. Rohrer, Study of ozone formation and transatlantic transport from biomass burning emissions over West Africa during the airborne Tropospheric Ozone Campaigns TROPOZ I and TROPOZ II, *J. Geophys. Res.*, *103*, 19059–19073, 1998.

Justice, C. O., J. D. Kendall, P. R. Dowty, and R. J. Scholes, Satellite remote sensing of fires during the SAFARI campaign using NOAA-advanced very high resolution radiometer data, *J. Geophys. Res.*, *101*, 23851–23863, 1996.

Kim, J.-H., and M. J. Newchurch, Climatology and trends of tropospheric ozone over the Eastern Pacific ocean, *Geophys. Res. Lett.*, *23*, 3,723–3,726, 1996.

Krishnamurti, T. N., M. C. Sinha, M. Kanamitsu, D. Oosterhof, H. Fuelberg, R. Chatfield, D. J. Jacob, and J. Logan, Passive tracer transport relevant to the TRACE-A experiment, *J. Geophys. Res.*, *101*, 23889–23907, 1996.

Lelieveld, J., P. J. Crutzen, D. Jacob, and A. M. Thompson, Modeling of biomass burning influences on tropospheric ozone, in B. W. van Wilgen (Ed.), *Fire in the Southern Africa*

Savannas: Ecological and Atmospheric Perspectives, University of Witwatersrand Press, Johannesburg, 1997, Chapter 10.

Levine, J. S., *Biomass Burning: Atmospheric. Climatic and Biospheric Implications*, MIT Press, Cambridge, MA, 1991.

Levine, J. S., *Biomass Burning and Global Change*, MIT Press, Cambridge, MA, 1996.

Liew, S. C., L. K. Kwo, K. Padmanabhan, O. K. Lim, and H. Lim, Delineating land/forest fire burnt scars with ERS interferometric synthetic aperture radar, *Geophys. Res. Lett.*, *26*, 2409–2412, 1999.

Longo, K. M., A. M. Thompson, V. W. J. H. Kirchhoff, L. A. Remer, S. R. de Freitas, M. A. F. S. Dias, P. Artaxo, W. Hart, J. D. Spinhirne, and M. A. Yamasoe, Correlation between smoke and tropospheric ozone concentration in Cuiabã during Smoke, Clouds, and Radiation-Brazil (SCAR-B), *J. Geophys. Res.*, *104*, 12113–12129, 1999.

Mauzerall, D. L., J. A. Logan, D. J. Jacob, B. E. Anderson, D. R. Blake, J. D. Bradshaw, B. Heikes, G. W. Sachse, H. Singh, and B. Talbot, Photochemistry in biomass burning plumes and implications for tropospheric ozone over the tropical South Atlantic, *J. Geophys. Res.*, *103*, 8401–8423, 1998.

Newell, R. E., V. Thouret, J. Y. N. Cho, P. Stoller, A. Marenco, and H. G. Smit, Ubiquity of quasi-horizontal layers in the troposphere, *Nature*, *398*, 316–319, 1999.

Olson, J. R., B. A. Baum, D. R. Cahoon, and J. H. Crawford, Frequency and distribution of forest, savanna and crop fires over tropical region during PEM-Tropics A, *J. Geophys. Res.*, *104*, 5865–5876, 1999.

Oltmans, S. J., A. S. Lefohn, H. E. Scheel, J. M. Harris, H. Levy, I. E. Galbally, E. G. Brunke, C. P. Meyer, J. A. Lathrop, B. J. Johnson, D. S. Shadwick, E. Cuevas, F. J. Schmidlin, D. W. Tarasick, H. Claude, J. B. Kerr, and O. Uchino, Trends of ozone in the troposphere, *Geophys. Res. Lett.*, *25*, 139–142, 1998.

Pickering, K. E., A. M. Thompson, Y. Wang, W-K Tao, D. P. McNamara, V. W. J. H. Kirchhoff, B. G. Heikes, G. W. Sachse, J. D. Bradshaw, G. L. Gregory, and D. R. Blake, Convective transport of biomass burning emissions over Brazil during TRACE-A, *J. Geophys. Res.*, *101*, 23993–24012, 1996.

Randriambelo, T., J. L. Baray, S. Baldy, P. Bremaud, and S. Cautenet, A case study of extreme tropospheric ozone contamination in the tropics using in-situ, satellite, and meteorological data, *Geophys. Res. Lett.*, *26*, 1287–1290, 1999.

Reichle, H. G., V. S. Connors, J. A. Holland, R. T. Sherrill, H. A. Wallio, J. C. Casas, E. P. Condon, B. B. Gormsen, and W. Seiler, The distribution of middle tropospheric carbon-monoxide during early October 1984, *J. Geophys. Res.*, *95*, 9845–9856, 1990.

Schultz, M. G., D. J. Jacob, Y. H. Wang, J. A. Logan, E. L. Atlas, D. R. Blake, N. J. Blake, J. D. Bradshaw, E. V. Browell, M. A. Fenn, F. Flocke, G. L. Gregory, B. G. Heikes, G. W. Sachse, S. T. Sandholm, R. E. Shetter, H. B. Singh, and R. W. Talbot, On the origins of tropospheric ozone and NO_x, over the tropical South Pacific, *J. Geophys. Res.*, *104*, 5829–5844, 1999.

Smit, H., D. Kley, S. McKeen, A. Volz, and S. Gilge, The latitudinal and vertical distribution of tropospheric ozone over the Atlantic Ocean in the southern and northern hemispheres, in R. D. Bojkov and P. Fabian (Eds.), *Ozone in the Atmosphere*, 1989, pp. 419–422.

Smyth, S. B., S. T. Sandholm, J. D. Bradshaw, R. W. Talbot, D. R. Blake, N. J. Blake, F. S. Rowland, H. B. Singh, G. L. Gregory, B. E. Anderson, G. W. Sachse, J. E. Collins, and A. S. Bachmeier, Factors influencing the upper free tropospheric distribution of reactive nitrogen

over the South Atlantic during the TRACE A experiment, *J. Geophys. Res.*, *101*, 24165–24186, 1996.

Swap, R. J., M. Garstang, S. A. Macko, P. D. Tyson, and P. Kållberg, Comparison of biomass burning emissions and biogenic emissions to the tropical south Atlantic, in J. S. Levine (Ed.), *Biomass Burning and Global Change*, MIT Press, Cambridge, MA, 1996, pp. 396–402.

Thompson, A. M., B. G. Doddridge, J. C. Witte, R. D. Hudson, W. T. Luke, J. E. Johnson, B. J. Johnson, and S. J. Oltmans, Shipboard and satellite views of elevated tropospheric ozone over the tropical Atlantic in January–February 1999, *Geophys. Res. Lett.*, *22*, 3317–3320, 2000.

Thompson, A. M., and R. D. Hudson, Tropical tropospheric ozone (TTO) maps from Nimbus 7 and Earth-Probe TOMS by the modified-residual method: Evaluation, El Niño signals and trends based on Atlantic regional time series, *J. Geophys. Res.*, 26961–26975, 1999.

Thompson, A. M., K. E. Pickering, D. P. McNamara, M. R. Schoeberl, R. D. Hudson, J. H. Kim, E. V. Browell, V. W. J. H. Kirchhoff, and D. Nganga, Where did tropospheric ozone over southern Africa and the tropical Atlantic come from in October 1992? Insights from TOMS, GTE/TRACE-A and SAFARI-92, *J. Geophys. Res.*, *101*, 24251–24278, 1996.

Thompson, A. M., W-K. Tao, K. E. Pickering, J. R. Scala, and J. Simpson, Tropical deep convection and ozone formation, *Bull. Am. Meteorol. Soc.*, *78*, 1043–1054, 1997.

Tyson, P. D., M. Garstang, A. M. Thompson, P. D'Abreton, R. D. Diab, and E. V. Browell, Atmospheric transport and photochemistry of ozone over central Southern Africa during the Southern Africa Fire-Atmosphere Research Initiative, *J. Geophys. Res.*, *102*, 10623–10635, 1997.

van Wilgen, B. W., M. O. Andreae, J. G. Goldammer, and J. A. Lindesay, *Fire in the Southern Africa Savannas: Ecological and Atmospheric Perspectives*, Witwatersand University Press, Johannesburg, South Africa, 1997.

Weller, R., R. Lilischkis, O. Schrems, R. Neuber, and S. Wessel, Vertical ozone distribution in the marine atmosphere over the central Atlantic Ocean (56°S–50°N), *J. Geophys. Res.*, *101*, 1387–1399, 1996.

Zenker, T., A. M. Thompson, D. P. McNamara, T. L. Kucsera, F. G. Wienhold, G. W. Harris, P. LeCanut, M. O. Andreae, and R. Koppmann, Regional trace gas distribution and airmass characteristics in the haze layer over southern Africa during the biomass burning season (Sep./Oct. 1992): Observations and modeling from the STARE/SAFARI-92/DC-3, in J. S. Levine (Ed.), *Biomass Burning and Global Change*, MIT Press, Cambridge, MA, 1996, pp. 296–308.

Zepp, R. G., W. L. Miller, R. A. Burke, D. A. B. Parsons, and M. C. Scholes, Effects of moisture and burning on soil-atmosphere exchange of trace carbon gases in a southern African savanna, *J. Geophys. Res.*, *101*, 23699–23706, 1996.

Ziemke, J. R., S. Chandra, and P. K. Bhartia, Two new methods for deriving tropospheric column ozone from TOMS measurements: The assimilated UARS MLS/HALOE and convective–cloud differential techniques, *J. Geophys. Res.*, *103*, 22115–22128, 1998.

Ziemke, J. R., S. Chandra, A. M. Thompson, and D. P. McNamara, Zonal asymmetries in southern hemisphere column ozone: Implication of biomass burning, *J. Geophys. Res.*, *101*, 14421–14427, 1996.

CHAPTER 15

ACID RAIN AND DEPOSITION

WILLIAM B. GRANT

1 INTRODUCTION

Early Concern

The first mention of acid rain in print was by Robert Boyle in which he referred to "nitrous or salino-sulphureous spirits" in air in his 1692 book *A General History of the Air*. The Scottish chemist Robert Angus Smith began to study acid rain in Manchester, England, in 1852 and extended the work in England, Scotland, and Germany for 20 years. His 1872 book, *Air and Rain: The Beginnings of a Chemical Climatology*, pointed out the link between sulfur pollution and "acid rain." He warned that acid rain was damaging plants and materials downwind of industrial regions, but his warning went largely unheeded.

While some research was conducted on acid deposition in the ensuing years, it was not until the 1950s and 1960s that E. Gorham, conducting research in England and Canada, built the major foundations for our present understanding of the causes of acid precipitation and its impact on aquatic ecosystems. However, it took the work of a Swedish scientist, S. Oden, in the 1960s, to arouse the scientific community and general public to engage in the debate about acid deposition. One newspaper account described his ideas about an insidious "chemical war" among the nations of Europe. Thus, by the 1970s, it was finally realized that Eastern Europe, Germany, Scandinavia, Canada, and the United States were experiencing widespread damage to forests and lakes as well as damage to stone and metal buildings and other structures from acid rain. In Germany, the term *Waldsterben* (forest death) was coined. Forests in parts of the Czech Republic, Slovakia, and Russia were practically devastated due to acid rain and heavy-metal ion deposition from uncontrolled industrial and power plant emissions. China and India are also experiencing significant effects of acid

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

rain, with the Taj Mahal losing much of its stonework surface material to acid deposition.

Acid Rain Chemistry

Acid rain is actually precipitation of various ions, both anions and cations, through precipitation, such as rain, snow, fog, as well as dry particles or aerosols. A typical ion balance is

$$[H^+] + [Na^+] + [Na_4{}^+] + 2[Ca^{2+}] = 2[SO_4{}^{2-}] + 2[SO_3{}^{2-}] + [NO_3^-] + [Cl^-] + [OH^-] + [HCO_3{}^-] + 2[CO_3{}^{2-}] \quad (1)$$

The primary naturally occurring trace gas that affects the pH of precipitation is carbon dioxide (CO_2), which forms carboxylic acid in water. The aqueous reactions of carbon dioxide are as follows:

$$CO_2 \text{ gas} + H_2O \rightarrow H_2CO_3 \quad (2)$$

$$H_2CO_3 \rightarrow HCO_3{}^- + H^+ \quad (3)$$

$$HCO_3{}^- \rightarrow CO_3{}^{2-} + H^+ \quad (4)$$

Since pK_a of (4) is as high as 10.3, reaction (2) has the greatest influence on the acidity of natural atmospheric systems. For a partial pressure of CO_2 of 350 ppmv, Henry's law constant (K_H) is as follows:

$$K_H = [H_2CO_3]/[CO_2 \text{ gas}] = 3.97 \times 10^{-2} \text{ mol/L atm} \quad (5)$$

and the equilibrium constant (K_3) of reaction (3) is given by:

$$K_3 = [H^+][HCO_3{}^-]/[H_2CO_3] = 4.5 \times 10^{-7} \text{ mol/L} \quad (6)$$

By combining and rearranging these two expressions, one arrives at the following equation:

$$[HCO_3{}^-] = ([CO_2 \text{ gas}] \times K_H \times K_3)/[H^+] \quad (7)$$

If the concentration of bicarbonate in water is equal to the hydrogen ion concentration, then by substitution, one arrives at the following:

$$[H^+]^2 = ([CO_2 \text{ gas}] \times K_H \times K_3)/[H^+] \quad (8)$$

$$= 5.97 \times 10^{-12} \text{ mol}^2/\text{L}^{-2} \quad (9)$$

Therefore,

$$[H^+] = 2.44 \times 10^{-6} \text{ mol/L} \tag{10}$$

and, hence

$$\text{pH} = -\log[H^+] = 5.6 \tag{11}$$

The bracketed quantities denote molar concentrations, with the cations on the left, the anions on the right.

Note that Henry's law can be expressed in terms of a pseudo-Henry's law constant to account for the increased uptake of gas in the liquid due to reactions in the liquid. For example,

$$K_{CO_2}^* = [CO_2 \times H_2O + HCO_3^- CO_3^{2-}]/P_{CO_2} \tag{12}$$

However, 5.6 is not necessarily the natural pH of rain since other naturally occurring species also play a role. Nitrogen oxides are formed naturally during lightning discharges, and sulfur species are released into the atmosphere over the oceans from biological activity as dimethyl sulfide (DMS). Hydrocarbon acids such as carboxylic acids, $HCOO_t$, and methylcarboxylic acids, CH_3COO_t, also contribute to the acidity, especially in remote, forested regions. On the other hand, base cations from soil dust, such as Ca, Mg, K, and P, etc., are alkaline and increase the pH. Thus, the natural acidity of precipitation can vary considerably depending on the upwind sources and, as will be discussed, meteorological conditions.

In addition, pollutants such as the nitrogen and sulfur oxides also contribute to acidity and are the focus of most of the concern regarding acid deposition. However, ammonia, often associated with agricultural operations, is alkaline. This chapter will examine the sources, the chemical transformations involved in the production of acid deposition, transport, deposition amounts and trends, and the effects on soils, plants, animals, and materials.

2 SOURCES

Natural

To put pollution contributions into perspective, it is worthwhile first to understand the role that naturally occurring materials play in acid deposition. Natural sources of sulfur account for 25 to 30% of the total, unless there are large volcanic eruptions, such as El Chichon in 1982 or Mount Pinatubo in 1991. Mount Pinatubo was estimated to emit 9 Tg of S into the stratosphere (total sources are 94 to 123 Tg S/yr), where the e^{-1} residence time for sulfuric acid aerosols is approximately one year.

Oxides of nitrogen ($NO_x = NO + NO_2$) are also produced naturally. As discussed in Chapter 4, natural sources such as soil emissions, lightning, stratospheric–tropospheric exchange, and a portion of biomass burning account for approximately one third of total NO_x.

Hydrocarbons are also involved in acid deposition. Carboxylic acids, $HCOO_t$, and methylcarboxylic acids, CH_3COO_t, are important hydrocarbon acids derived from direct terrestrial emissions as well as oxidation of emissions by marine or terrestrial biota.

Base cations are generally derived from soils through lofting of aeolian dust by wind. Deserts, such as the Sahara Desert in Africa and the Gobi Desert in China, generate large dust clouds each year that are transported thousands of kilometers. The dust from the Sahara Desert often reaches both North and South America and may provide significant base cations for vegetation in the rain forests. The dust from the Gobi Desert is often seen over Japan and the Korean Peninsula. The base cation deposition in Europe was studied for 1989. Using a 10×20 km^2 grid, maps were produced showing that base cations neutralize $SO_4^{2-} + NO_3^-$ by much more in southern regions than in northern regions. South of 45° to 50°N, more than 50% was neutralized, with more than 75% in some locations; in Norway and Sweden, the amount neutralized generally fell to less than 10%. The variations can be explained in terms of the amounts of acid ions and base cations in the air. Soil-derived dust in the United States used to provide the base cations to help neutralize the effects of sulfur and nitrogen. However, the amount of base cations in precipitation has been declining in the past 2 to 3 decades in the United States probably because of changes in farming and construction practices that leave fewer disturbed regions from which wind can raise dust.

Anthropogenic

Anthropogenic sources of sulfur accounted for 77% of global sulfur emissions in 1980. Combustion of fossil fuel for electric power production is responsible for most of the anthropogenic contributions to acid deposition, accounting for 67% of the anthropogenic SO_2 emissions in the United States in 1996. Industrial fuel combustion accounts for 17% of the U.S. SO_2 emissions, with various other sources accounting for the rest. Fuel used for transportation generates 7% of the SO_2 emissions, which has been linked to regional haze patterns in such places as the Los Angeles Basin and portions of the eastern United States.

NO is a by-product of combustion of all hydrocarbon fuels, both fossil fuel and fresh biomass, due to the high temperatures involved. In the United States in 1996, 30% of the NO emissions came from on-road vehicles, 28% from electric utilities, 19% from nonroad engines and vehicles, 13% from industrial fuel combustion, and 10% from other sources. On a global basis, anthropogenic NO emissions are highest where industry, fossil fuel power plants, and surface transportation are most densely sited, i.e., the northern midlatitudes.

Anthropogenic ammonia emissions are associated with fertilizers and livestock feedlots. Organic acids also contribute to the anthropogenic burden of acid deposi-

tion. The major organic acids found in the gas phase are formic acid (HCOOH) and acetic acid (CH_3COOH), with other organic acids found in minor amounts. Sources include automobile exhaust, biomass burning, and some food processing plants.

3 TRANSFORMATION

The emitted SO_2, NO, and NH_3 are transformed to aerosols and components of precipitation through both gas-phase and aqueous-phase chemical reactions.

Sulfur dioxide is transformed in the gas phase primarily by:

$$SO_2 + OH \rightarrow HOSO_2 \tag{13}$$

Ozone can also lead to the oxidization of SO_2. Such reactions would be especially important at night when OH radical concentrations are very small due to the absence of solar radiation. One way this can happen is for ozone to react with an alkene, such as ethene or propene by adding to the carbon double bond, creating a primary ozonide. Since ozonides are not stable, this can rapidly split into what is called a Criegee intermediate, named after the German chemist who proposed the mechanism. A Criegee intermediate can react with SO_2 in a series of steps that also result in the oxidation of SO_2, which can also be oxidized directly by ozone, but the reaction rate is slow. Note that the rate of oxidation of SO_2 has a seasonal cycle in middle latitudes, being as much as an order of magnitude lower in winter than in summer.

In the aqueous phase, other reactions can occur. For example:

$$SO_2 + H_2O \rightarrow SO_2 \times H_2O \tag{14}$$

$$SO_2 \times H_2O \rightarrow HSO_3^- + H^+ \tag{15}$$

$$HSO_3^- \rightarrow SO_3^{2-} + H^+ \tag{16}$$

These reactions establish equilibria of the various sulfur species, with mole fractions dependent on the pH of the solution and both Henry's law constant [for (14)] and equilibrium constants [for (15) and (16)]. Dissolved SO_2 (13) is favored at pH below 2, the bisulfite ion (15) for $2 < pH < 7$, and the sulfite ion (16) for $pH > 7$.

Aqueous-phase reactions with H_2O_2 in cloud, fog, and raindrops are considered to be the dominant mechanisms for the oxidation of SO_2 to H_2SO_4. Thus, H_2O_2 could be rate limiting. Field and modeling studies show that to explain the seasonal concentrations of H_2O_2 (higher in summer than in winter) the initial rate of aqueous phase H_2O_2 photoformation has to be linearly dependent on solar actinic flux, i.e., radiation that induces photochemical reactions. Organic chromophores are suggested to be responsible for the H_2O_2 photoformation. One implication of this study is that the seasonal variability in the nonlinearity between SO_2 emissions and regional sulfate deposition may be largely explained. Other peroxides can also oxidize SO_2, but exist in lower concentrations than does H_2O_2.

Other reactions leading to the oxidation of S(IV) include ozone, O_2 catalyzed by transition metal ions such as Fe^{3+} and Mn^{2+}, and carbonaceous particles. While the reactions with H_2O_2 are generally most important (2 to 20% per hour, independent of pH), the others are much weaker in general and have very strong pH dependences. Above a pH of 5, the reaction with ozone is comparable to that for H_2O_2, with the other reactions somewhat weaker.

Note that other sulfur species, such as hydrogen sulfide (H_2S) and carbonyl sulfide (OCS), emitted by biological sources, can also be oxidized, as well as dimethyl sulfide (DMS), emitted from marine sources. While OH is the primary source of DMS oxidation, NO_3 also reacts rapidly with DMS, and halogens, such as bromine, chlorine, and iodine are also potential reactants with DMS in the marine boundary layer.

Nitric oxide (NO) is rapidly oxidized to NO_2, especially by reacting with ozone:

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{17}$$

From there, it is transformed to nitric acid by interaction with the hydroxyl radical:

$$NO_2 + OH \rightarrow HNO_3 \tag{18}$$

This reaction is about 10 times more rapid than that of (13).

Nitric acid can also be formed by the reaction with various organics, such as the alkanes and aldehydes. In this case, hydrogen is abstracted from the organic molecule. This reaction may account for 15% of the nitric acid formation, occurring primarily at night.

Both sulfate and nitrate aerosols are very hygroscopic and increase in diameter rapidly with increases in relative humidity above 50% to 70%. In the absence of cloud formation, they form the bulk of regional aerosols downwind of heavily industrialized/urbanized regions, such as the eastern United States. As acid haze becomes thicker and stays near the surface, it can become acid fog, such as has been observed in California and in eastern U.S. mountains. An aerosol/fog cycle can be set up in which aerosol particles grow by water condensation on existing nuclei, dilute and dissolve in fog droplets, where they undergo chemical conversions. The process can go the other way as solute concentrations increase due to evaporation of the water, leading back to aerosols. Thus, as the temperature cycles during the day, the fog–aerosol–fog cycle can be made.

An excellent overview of the chemistry of acid precipitation can be found in Finlayson-Pitts and Pitts (2000).

4 TRANSPORT

In addition to source regions and transformation mechanisms and rates, winds and other meteorological conditions also play important roles in determining where acid precipitation will occur. The pollution plumes will be transported at the rate of the

prevailing winds. The source gases will be transformed at various rates depending on such factors as amount of solar radiation, concentrations of OH and water vapor, temperature, and the extent of clouds.

Sulfate can be transported up to 1100 km in normal downwind directions and up to 400 km in the normal upwind directions (i.e., during the reduced opportunities for transport in that direction), while nitrogen oxides are transported as nitrates as far as 200 to 800 km. It is found that turnover times for anthropogenic sulfate are 4.7 ± 1.1 days in the eastern United States.

The transport of ammonia and ammonium depends on the emissions of SO_2 and NO_x along the trajectory of the air mass containing them. The transport distance for ammonia and ammonium in northern Europe depends on the amount of SO_2 and NO_x present. When they are present, transport is reduced significantly because ammonium aerosols are formed rapidly. NH_x is most likely to be deposited in the country of origin in Europe, given the sizes of the countries, while for SO_x only 25 to 30% would be deposited, and for NO_x only 10%.

5 DEPOSITION

Acid deposition occurs in two primary forms—wet and dry. Wet deposition comprises rain, snow, and fog. Dry deposition involves turbulent transport of aerosol and gases to the surface layer. The relative amounts of wet and dry deposition depend on a number of factors, such as the amount of precipitation, whether the

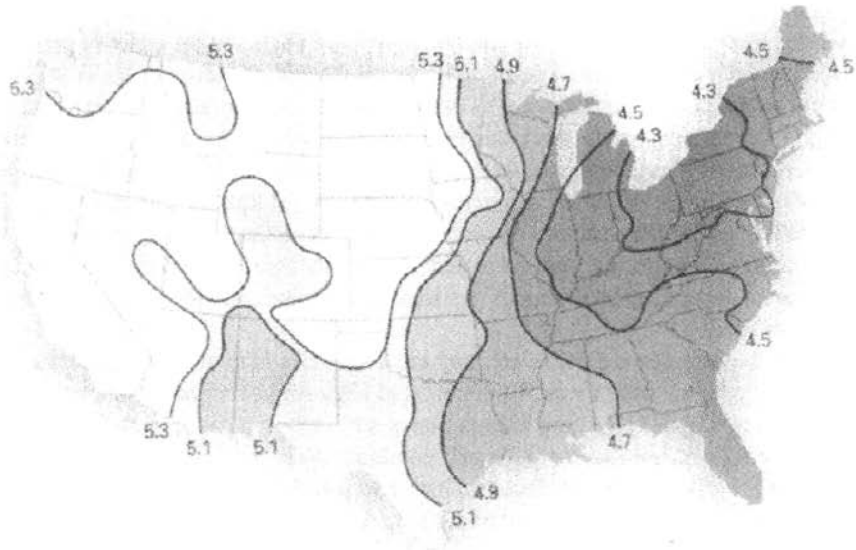


Figure 1 Annual pH of rain for the United States in 1990. The black lines indicate contours of equal pH. See ftp site for color image.

elevation is above the cloud line, how far the site is from the primary sources of the acid ions, etc. At U.S. Environmental Protection Agency (EPA) National Dry Deposition Network stations in the eastern United States in 1991, dry sulfate deposition accounted for approximately 10 to 60% (mean approximately 40%) of total sulfate deposition, with wet deposition accounting for the rest. For nitrates, the dry deposition fraction varied from 20 to 65% (mean approximately 45%). Due to the seasonal cycle in the rate of oxidation of SO_2, deposition rates for SO_2 tend to be higher than for SO_4 in winter, with the reverse occurring in summer.

The acidity of deposition depends on the difference between anions and cations in the precipitate. Thus, the nitrate and sulfate ions reduce the pH, while ammonium and soil-derived dust increase the pH. Figure 1 shows a map of the pH of rain for the United States in 1990, indicating that the pH is lowest just southeast of the Great Lakes, a consequence of the high amount of fossil fuel combustion in and to the west of the region.

6 MEASUREMENT

Instruments

Various instruments are used in the study of acid deposition. Since emission rates are generally estimated based on factors associated with fuel consumption, not many measurements are made at the source regions. Standard meteorological instruments and networks are used for the meteorological data input. The collectors generally use polypropylene funnels and bottles. The bottles may be refrigerated to 4°C to reduce evaporation and/or heated to melt snow. When wet and dry deposition collectors are used together, a lid is placed over the dry deposition bucket at the onset of precipitation, then back over the wet deposition bucket at the end of precipitation. However, it should be noted that measurement of dry deposition is notoriously difficult, and that buckets do not adequately represent the manner in which the local surfaces collect dry deposition. The three conceptual ways in which dry deposition is measured are: (1) direct collection on surrogate or natural surfaces, (2) flux measurements by eddy correlation or profile techniques, and (3) indirect estimation using atmospheric concentration monitoring and estimated deposition velocities. Which approach is used varies depending on the funds available and the accuracy to which the information is desired.

Once the samples are collected, they are taken to a laboratory for analysis. The analytical methods used by the National Acid Deposition Program/National Trends Network (NADP/NTN) in the United States are likely typical of such programs. A glass electrode is used to measure pH; conductivity is measured using a platinum electrode; chloride, nitrate, orthophosphate and sulfate are measured with ion chromatrography with a detection limit of 0.03 mg/L for all but orthophosphate, which is measured with a detection limit of 0.02 mg/L; ammonium is measured using automated phenate colorimetry with a detection limit of 0.02 mg/L; calcium, magnesium, potassium, and sodium are measured with flame atomic absorption spectro-

photometer with a detection limit of 0.003 except for calcium, for which the detection limit is 0.09 mg/L. Sodium and/or magnesium can be used to estimate the fraction of material derived from sea salt. This is useful in determining how to apportion the sulfate values between terrestrial and oceanic sources.

The Acid Precipitation in Ontario Study (APIOS) deposition monitoring program has similar instrumentation with slightly different detection limits. The NADP/NTN detection limits were improved by instrument changes in 1985, while the APIOS instrumentation was established in 1980 and not updated as of 1990.

Surface networks

Collection instruments are often set out in networks. The NADP/NTN is an example of such a network. It is part of a cooperative program that includes federal, state, and private research organizations. The objectives of the program are:

1. To measure and characterize the supply of beneficial and injurious chemical substances in atmospheric deposition on a broad regional scale
2. To determine the spatial patterns and temporal trends in the distribution of chemical elements deposited on natural and managed ecosystems
3. To provide information needed to gain a better understanding of the sources, transport, and transformation of materials contributing to or associated with acidic atmospheric deposition in the United States

The NADP/NTN was made operational in July 1978 and continues to the present time. The sites were selected to represent major physiographic, agricultural, aquatic, and forested areas throughout the United States. In general, sites are located in rural areas away from sources that could affect the measurements. The program grew from 22 sites in late 1978 to about 200 sites in 1985, which were still in operation in 1990. The containers are heated to 4°C to melt snow but are not refrigerated. Samples are collected weekly and sent to the Central Analytical Laboratory in Champaign, Illinois.

7 INTENSIVE STUDY PROGRAMS

In the 1980s, a major study, the National Acid Precipitation Assessment Program (NAPAP) was funded by Congress to investigate the situation in the United States. The total cost was $500 million. Areas of investigation included acid deposition and effects on aquatic and terrestrial ecosystems. Both nitrate and sulfate depositions were found to be highest in the northeast United States near the eastern Great Lakes, centered on eastern Michigan, western New York and Pennsylvania, and northern West Virginia, and extending into southern Ontario, albeit with somewhat different geographical distributions. Ammonium deposition peaked in Michigan and southern

Ontario. As a consequence, annual pH of rain is lowest in New York, Pennsylvania, and West Virginia as shown in Figure 1 for 1990.

Similar programs have been carried out in a number of European countries, especially in terms of acid rain effects on forests, with a number of them reported in the Springer *Ecological Studies* series.

8 GLOBAL TRENDS IN EMISSIONS AND DEPOSITION

With accelerating economic development in Southeast Asia, anthropogenic NO_x emissions are expected to increase dramatically in the near future. It has been estimated that global NO_x emissions will increase from an estimated 19 Tg NO_2 in 1990 to 86 Tg NO_2 in 2020. The largest increases are expected in the power and transport sectors.

Trends of acid deposition should generally follow the regional trends for fossil fuel consumption, with coal and oil providing most of the sulfur, and all components contributing to the nitrogen oxides and organic acids. In the United States, wood was the primary source of fuel until 1880, being used to generate about 3×10^{15} Btu/yr at the peak in 1870. Coal started to be used in increasing amounts around 1850, rising to 15×10^{15} Btu/yr by 1916, staying in the range 10 to 17×10^{15} Btu/yr after that. Oil started to become an important fuel source after 1900, rising to 35×10^{15} Btu/yr by 1977 before leveling off. Natural gas also became important after 1900, rising to 24×10^{15} Btu/yr by 1970 before dropping slightly. Thus, in the United States, acid deposition should have risen steadily from 1900 to at least the 1980s. In the eastern and midwestern United States there has been an estimated 19% decrease in SO_2 emissions and a 16% decrease in NO_x emissions between 1975 and 1987. Since the U.S. Clean Air Act Amendments of 1990 mandated further decreases in sulfur emissions, they have continued to decrease. Between 1989 and 1995, sulfur dioxide decreased 35% and sulfate 26% in rural eastern United States. Nitrogen emissions have not been recognized as being very important until recently for a variety of scientific and political reasons, and it is more difficult to remove NO_x than SO_2 from the flue gases, so the regulations on nitrogen emissions are not as strong as for sulfur. Between 1989 and 1995 nitrogen concentrations in rural eastern United States had fallen only 8%.

Data for historical anthropogenic emissions of SO_2 are also available for Europe. A gradual increase is seen from 1880 (0.45 Tg/yr) to 1940 (1.4 Tg/yr), a dip in 1945, then a rapid increase to >36 Tg/yr in 1980, followed by a gradual decline thereafter. Ammonia emissions peaked in the mid-1980s.

Continued population growth and development are expected to lead to an increase of 25% in the deposition of nitrogen in the more-developed-country regions by the year 2020. Earth's population is projected to increase from 6 billion in 1999 to 8.5 billion in 2020, and per-capita energy consumption is expected to double compared to 1980. Much of the increase will be felt in Asia. The increases in nitrogen oxides may lead to larger ozone concentrations, thereby increasing the

oxidizing capacity of the atmosphere and its ability to absorb thermal infrared radiation.

9 SOIL CHANGES

Bernhard Ulrich is credited with determining how acid deposition affects soil during the acidification process. His 1966 study set the stage for his later work. His review summarizes the effects of acid deposition on soil cation-anion budgets and lists a number of his key works. As soil acidity increases due to acid deposition (or plant biomass harvesting for that matter), the base cations (e.g., Ca, Mg, K, P) try to neutralize the acidity and are leached from the upper soil horizons in the process. As the process continues, the transition metal and aluminum oxides are dissolved, with these cations becoming more prevalent in the soil solution. Nitric acid is a stronger acid than sulfuric, so it has a greater ability to lower the soil pH. An interesting recent finding is that as the process continues, Al^{3+} seems to accelerate the base cation leaching process, making Al^{3+} more readily available. As acid deposition continues over a long period, the acid neutralizing capacity (ANC) or alkalinity decreases.

Additional influences on ANC arise from biogeochemical processes. Trees, for example, enhance the collection of dry deposition as well as remove base cations from the soil. Soil organic matter storage is followed by decay, which releases the trace minerals, nitrogen, and organic acids. In addition, forest defoliation by the gypsy moth has exacerbated the effects of acidic deposition. Changes in stream water composition following severe defoliation of forested mountain watersheds in western Virginia has included increased concentrations of nitrate and acidity, as well as accelerated export of base cations, and pH and ANC reached lower levels than previously observed, especially during storm flow conditions. To date, several years following the defoliation, stream water composition has not returned to pre-defoliation values.

Finally, there are interactions between the various processes. Changing acid–base status changes vegetation amounts and types. Reductions in vegetation cover can lead to reduction in enhanced collection of dry deposition as well as higher surface temperatures, thereby increasing microbial activities.

10 EFFECTS ON FORESTS, AQUATIC ECOSYSTEMS, AND MATERIALS

Forests

Paradoxically, one of the first effects of acid deposition on trees and forests is that of stimulating growth, rather than hindering it. Nitrogen in both ammonium (NH_4) and nitrate (NO_3) forms can be utilized by trees in building amino acids required for growth. Thus, nitrogen deposition first has the impact of fertilizing plants. This

process eventually ceases in temperate ecosystems when the soil is nitrogen saturated. The impact of nitrogen deposition on carbon uptake by terrestrial ecosystems has been modeled using several different three-dimensional models. Both NO_y and NH_x deposition were considered. The bulk of the NO_y deposition was found to be in the eastern United States, Europe, and, to a lesser extent, in eastern Asia and Japan. All five models predict that most of the carbon will be sequestered in the forests of eastern United States and Europe. Without N saturation, C sequesterization was found to range from 6 to 13×10^{15} g C/yr, while with N saturation, the range was 5 to 10×10^{15} g C/yr. This implies that N saturation reduces the growth rate of forests, in line with what has been observed in forests in the northeastern United States.

Acid deposition also causes the soil solution pH to be lowered, in part through the increased biomass growth rate, since the tree has to give up hydronium ions in exchange for base cations. It should be noted that the impact of acid deposition on forests is mediated through the soils, with some better able to buffer the acid than others. Calcium carbonate or limestone, for example, has a high buffering capacity, and would take a long time to show serious effects from acid deposition. One response of trees is for the tree roots to try to grow away from the acid soil, which may take the form of growing more in the upper organic layer, rather than in the lower mineral horizons. This makes trees susceptible to other stresses, such as winds and drought. Another effect is that since trees obtain less calcium after long-term acid deposition, the strength of the boles (trunks) and branches is reduced, since plants rely upon calcium for cell wall structure, they are much more susceptible to falling during ice, snow, and wind storms, as was the case in the northeastern United States and southeastern Canada in early 1998.

Starting around the 1970s, researchers in the United States and Europe began to notice that trees were beginning to show evidence of decline for nonhistorical reasons. Acid deposition was identified as a likely suspect in the early 1970s, although the effects of acid deposition had been observed in the sixteenth century in Europe and discussed again in the midnineteenth century.

The effects of acid precipitation on European forests in the 1980s have been well documented, especially to the Norway spruce [*Picea abies* (*L.*) *Karst*]. A study investigating the spruce decline determined that a long history of acid deposition, mostly sulfate prior to the early part of the century, with nitrate added around 1915, led to the observed effects. The soils were somewhat deficient in calcium and magnesium, and by about 1980, there was a strong nutritional imbalance due to years of ammonium nitrate depositions, nitrate leaching from the soils, and soil acidification. The yellowing of the leaves was attributed to deficiencies in magnesium. While *Waldsterben* in Europe was less pronounced in the early-to-mid-1990s than in the mid-1980s, probably due to reductions in sulfur emissions, declines in forest health are still quite prevalent, especially in central Europe. Annual forest condition surveys in conjunction with the modeling studies of nitrogen deposition show increased soil acidity in the regions with highest forest decline symptoms. There, the mean plot defoliation was in the range of 20 to 40% in 1997, with evergreens affected more than deciduous trees.

Acid deposition has had an adverse impact on forests in the eastern United States. The decline of the red spruce forests in the northeastern United States has been attributed to acid deposition, as has the decline of red spruce forests in North Carolina. Acid deposition has also adversely affected the sugar maples (*Acer saccharum Marsh.*) in Pennsylvania and Quebec as well as red oaks (*Quercus rubra*) and white oaks (*Quercus alba*) in the eastern United States. Evidence linking acid deposition to U.S. forest condition is found using the U.S. Department of Agriculture Forest Service Forest Inventory and Analysis data in conjunction with acid ion deposition doses using the acid deposition data from NAPAP. Increased mortality rates for white oaks (*Quercus alba*) in the northeastern United States can be related statistically to increased acid ion doses.

Further evidence for the role of acid deposition affecting oaks is found in oak tree ring studies in North Carolina and Missouri in the United States. The growth spurts in the 1950s for oaks in decline compared with lower growth rates of healthier nearby oaks are consistent with the N fertilization effect; the gradual growth decline subsequently is consistent with impaired tree vitality due to both acid deposition and ozone exposure; and the rapid decline after major droughts in the 1980s is consistent with shallower root depth, leading to greater water stress in drought periods.

Aquatic Ecosystems

Aquatic ecosystems have borne much of the brunt of acid deposition, resulting in significant loss of invertebrate populations and fish production among other things.

There are several processes influencing acid–base chemistry of surface waters. Wet and dry deposition is one. The other important factor is the ANC (alkalinity) of the water body. In turn, the ANC is strongly affected by the soils and bedrock under and near the body. The difference between the sums of base cations and acid anions derived from the soils and bedrock is equal to the ANC. Location of a body of water in a region where the soils and rocks are more likely to contribute base cations than acid anions to the water are less likely to be acidified by acid deposition. The base cations involved at the higher ANC levels are generally, in approximate order of importance, calcium, magnesium, sodium, and potassium. The acid anions are, likewise, carbonate, organic acids, sulfate, chloride, and nitrate. Of course, local conditions affect the amounts and relative orders.

Both aquatic animals and plants are adversely affected by acidification. The processes affected by acidification include change rates and amounts of primary production, nutrient cycling, and decomposition. Aluminum plays an important role in acidified systems since it is detrimental or toxic to both animal and plant life. Normally, aluminum is tightly bound to oxygen or the hydroxyl radical, OH. As the pH is lowered below 6, the concentration of monomeric aluminum rises rapidly. Aluminum in acidified streams has been found to coat the gills of fish, leading to premature mortality.

It has recently been recognized that atmospheric deposition of nitrogen is playing a significant role in the eutrophication in estuaries and coastal waters, such as the Chesapeake Bay in the mid-Atlantic eastern United States. Until a landmark study

was published in 1991, it was thought that most of the nitrogen reaching such bodies of water came from agricultural operations. More recent work has determined that approximately 20% of the nitrogen reaching the Chesapeake Bay as wet precipitation is in the form of dissolved organic nitrogen. In addition, a significant fraction comes from ammonium.

Another consequence of lake acidification is increased transparency. Most likely this is due to reduction in dissolved organic carbon or from a change in the chemical nature and light absorption capacity of dissolved organics in the water. This can lead to changes in primary productivity and thermal structure at lower depths. An additional consequence of increased transparency is increased transmission of ultraviolet B (UV-B) (280 to 320 nm) radiation. This leads to reductions in abundances of phytoplankton and zooplankton sensitive to UV-B.

The geographic overview of the regional case study areas is instructive. The key factors distinguishing among the regions are geology, soils, climate, hydrology, deposition chemistry, land use, vegetation, and landforms. All play important roles in determining the degree of acidification of aquatic ecosystems. Regions with bedrock highly resistant to chemical weathering are more likely to have low ANC lakes. Calcareous bedrock leads to high ANC waters. However, if the overlying till has different properties, it can counter the influences of the bedrock. In the northeast United States, glaciers brought in till from the calcareous Canadian Shield, leading to high ANC lakes. Among soils, the younger soils, more often found in the northern United States, lead to lower ANC water bodies, while the older soils, more often found in the southeastern United States, lead to higher ANC water bodies due to the accumulated organic matter that can lead to organic acidity.

Materials

Acid deposition also affects materials such as rocks and metals used in monuments and building construction through corrosion. Calcareous rock materials such as marble and sandstone are particularly vulnerable since the base cations are leached by the acids just as in soils. Mortar from limestone is also very susceptible to damage, but bricks are largely immune to the effects. Even ancient monuments are affected in a variety of ways including removal of material; development of rusty yellow patinas rich in Fe and Cu; firmly attached black crusts in contact with percolating water, where recrystallized calcite shields amorphous deposits rich in S, Si, Fe, and carbonaceous particles; and black loose deposits of gypsum and fly ash particles. Also, metals that react with hydrogen, nitrate, or sulfate, such as copper and iron, will be slowly eroded. Modern building practices have to consider effects of acid deposition and corrosion in the design phase.

11 POLICIES

Since acid rain has adverse impacts on animals, plants, and structures, there is concern that levels be reduced from current levels in many places and not increase rapidly in developing regions where fossil fuel combustion is increasing.

After completion of the NAPAP study, but not because of it, the 1990 amendments to the Clean Air Act mandated reductions in sulfur dioxide emissions from power plants in an effort to reduce the impacts of acid deposition on the environment. The key study in this regard was one published in *Science* showing essentially that what goes up must come down, i.e., that regions within a few hundred miles downwind of SO_2 (and NO_x) emission sources would be impacted by the emissions.

Given the fact that anthropogenic emissions of acid precursors are expected to rise, and that acid deposition has major adverse impacts on both aquatic and land ecosystems, it seems to be worthwhile to set local, national, and international policies that would tend to reduce the projected increases in emissions. The four main routes to cutting pollution emissions are: (1) using low-pollutant fuels, (2) preventing the formation of pollutants such as NO during combustion, (3) screening pollutants from exhaust and flue gases, and (4) energy conservation. Some of these routes would also help reduce the emissions of greenhouse gases. Choosing between these routes or some combination thereof involves consideration of the trade-offs including economic and political issues, e.g., the sources of the various fuels and whether the costs of emissions reductions outweigh the benefits, with the added complication that the groups incurring the costs are not necessarily the ones reaping the benefits.

A variety of policies has been identified that could be adopted to reduce the contribution of transport sector NO_x emissions at the local level in the Netherlands. The most important national policies identified relate to vehicles and fuels, pricing policy, public transport policies, and national guidelines for policies on parking and land use, while the most important local policies identified are those for parking, land use, cycling, and restrictions for motorized vehicles.

Regulations that would lead to further reductions in nitric oxide and sulfur emissions were proposed in the United States in late 1999. Oil refiners are being asked to remove 90% of the sulfur from gasoline. The manufacturers of sport utility vehicles (SUVs) and light-duty trucks are being asked to comply with the emission standards for passenger vehicles. Older electric power generating plants, which tried to escape emissions controls under the "grandfather" clause, are being asked to cut their nitrogen emissions. The proposed action affects 392 generating units at both electric generating (EGU) and non-electric-generating (non-EGU) facilities in 12 states. Affected EGUs will be required to reduce NO_x emissions to $0.15\,\text{lb}\,(\text{mm Btu})^{-1}$, while large non-EGUs will be required to reduce NO_x emissions by approximately 60% from baseline levels.

If changed regulations are not sufficient, Congress may consider additional legislation to reduce emissions. Of course, there would be a phase-in period, so it might take a decade or two for the changes to have an impact on the environment.

REFERENCES

Adriano, D. C., and A. H. Johnson (Eds.), *Acidic Precipitation*, Vol. 2: *Biological and Ecological Effects*, Springer-Verlag, Berlin, 1989.

Boyle, Robert, *The General History of the Air*, Awnsham and John Churchill, London, 1692.

Charles, D. F., and S. Christie (Eds.), *Acidic Deposition and Aquatic Ecosystems*, Springer-Verlag, Berlin, 1991.

Cowling, E. B., Acid precipitation in historical perspective, *Environ. Sci. Technol.*, *16*, 110A–123A, 1982.

Erisman, J. W., and G. P. J. Draaijers, *Atmospheric Deposition in Relation to Acidification and Eutrophication*, Elsevier, New York, 1995.

Finlayson-Pitts, B. J., and J. N. Pitts, Jr., Acid deposition: formation and fates of inorganic and organic acids in the troposphere, Ch. 8 in *Chemistry of the Upper and Lower Atmosphere: Theory, Experiments and Applications*. Academic, New York, 2000, pp. 294–348.

Fisher, D. C., and M. Oppenheimer, Atmospheric nitrogen deposition to the Chesapeake Bay Estuary, *Ambio*, *23*, 102–108, 1991.

Graedel, T. E., and R. McGill, Degradation of materials in the atmosphere, *Environ. Sci. Technol.*, *20*, 1093–1100, 1986.

Hedin, L. O., and G. E. Likens, Atmospheric dust and acid rain, *Sci. Am.*, *275*(6), 88–92, 1996.

Johnson, D. W., and S. E. Lindberg (Eds.), *Atmospheric Deposition and Forest Nutrient Cycling*, Springer-Verlag, Berlin, 1992.

Likens, G. E., and F. H. Bormann, Acid rain: A serious regional environmental problem, *Science*, *184*, 1176–1179, 1974.

Lindberg, S. E., A. L. Page, and S. A. Norton (Eds.), *Acidic Precipitation*, Vol. 3: *Sources, Deposition and Canopy Interactions*, Springer-Verlag, Berlin, 1990.

Radojevic, M., and R. M. Harrison (Eds.), *Atmospheric Acidity, Sources, Consequences and Abatement*, Elsevier Applied Science, New York, 1992.

Schulze, E.-D., O. L. Lange, and R. Oren (Eds.), *Forest Decline and Air Pollution, Ecological Studies 77*, Springer-Verlag, Berlin, 1989.

Schütt, P., and E. B. Cowling, Waldsterben, a general decline of forests in Central Europe: Symptoms, development, and possible causes, *Plant Disease*, *69*, 548–558, 1985.

Schwartz, S. E., Acid deposition: Unraveling a regional phenomenon, *Science*, *243*, 753–763, 1989.

Sisterson, D. L., V. C. Bowersox, T. P. Meyers, A. R. Olsen, and R. J. Vong, *Deposition Monitoring: Methods and Results*, NAPAP Report 6, Argonne National Laboratory, Argonne, III, 1990.

Smith, Robert A., *Air and rain [microform]: The Beginning of a Chemical Climatology*, Longmans, London, 1872.

Sverdrup, H., and P. Warfvinge, Past and future changes in soil acidity and implications for forest growth under deposition scenarios, *Ecol. Bull.*, *44*, 335–351, 1995.

Ulrich, B., Nutrient and acid–base budget of Central European forest ecosystems, in *Effects of Acid Rain on Forest Processes*, Wiley-Liss, New York, 1994, pp. 1–50.

Ulrich, B. 1983(a). A concept of forest ecosystem stability and of acid deposition as a driving force for destabilization. In: Ulrich, B. and Pankrath, J (Eds), Effectsm of Accumulation of Air Polutants in Forest Ecosystems. D Reidel Publishing Company, 1–29.

Ulrich, B. 1983(b). Soil acidity and its relations to acid deposition. In: Ulrich, B and Pankrath, J (Eds), Effects of Accumulation of Air Pollutants in Forest Ecosystems. D Reidel Publishing Company, 127–146.

CHAPTER 16

FUNDAMENTALS OF VISIBILITY

WILLIAM C. MALM

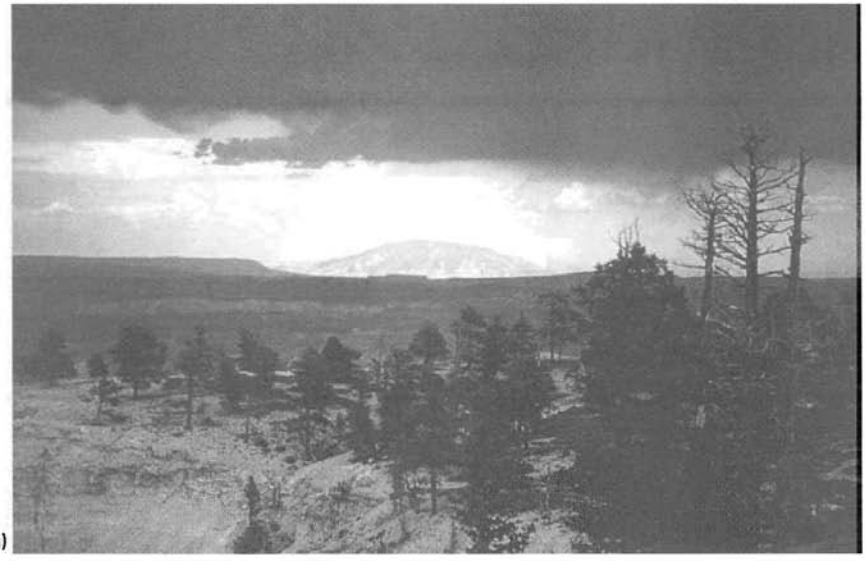
1 INTRODUCTION

A definition of visibility, as it relates to management of the many visual resources found in national parks, wilderness areas, and urban centers, is a complex and difficult concept to address. Should visibility be defined in strictly technical terms that concern themselves with exact measurements of illumination, threshold contrast, and precisely measured distances? Or is visibility more closely allied with value judgments of an observer viewing a scenic vista?

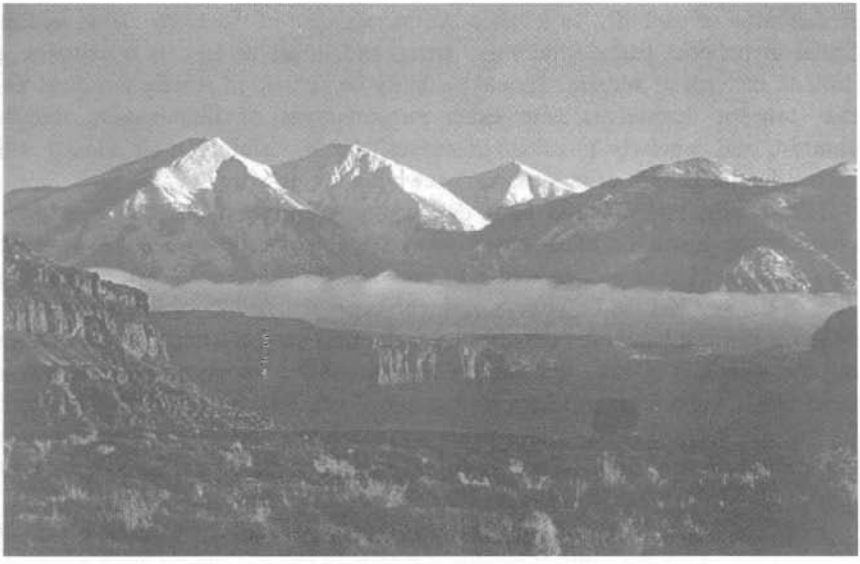
Historically, *visibility* has been defined as the greatest distance at which an observer can just see a black object viewed against the horizon sky. An object is usually referred to as at threshold contrast when the difference between the brightness of the sky and the brightness of the object is reduced to such a degree that an observer can just barely see the object. Much effort has been expended in establishing the threshold contrast for various targets under a variety of illumination and atmospheric conditions. An important result of this work is that threshold contrast for the eye, adapted to daylight, changes very little with background brightness, but it is strongly dependent upon the size of the target and the time spent looking for the target.

However, visibility is really more than being able to see a black object at a distance for which the contrast reaches a threshold value. Coming upon a mountain such as one of those shown in Figures 1*a* and 1*b*, an observer does not ask, "How far do I have to back away before the vista disappears?" Rather, the observer will comment on the color of the mountain, on whether geological features can be seen and appreciated, or on the amount of snow cover resulting from a recent storm system. Approaching landscape features such as those shown in Figures 1*c*

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

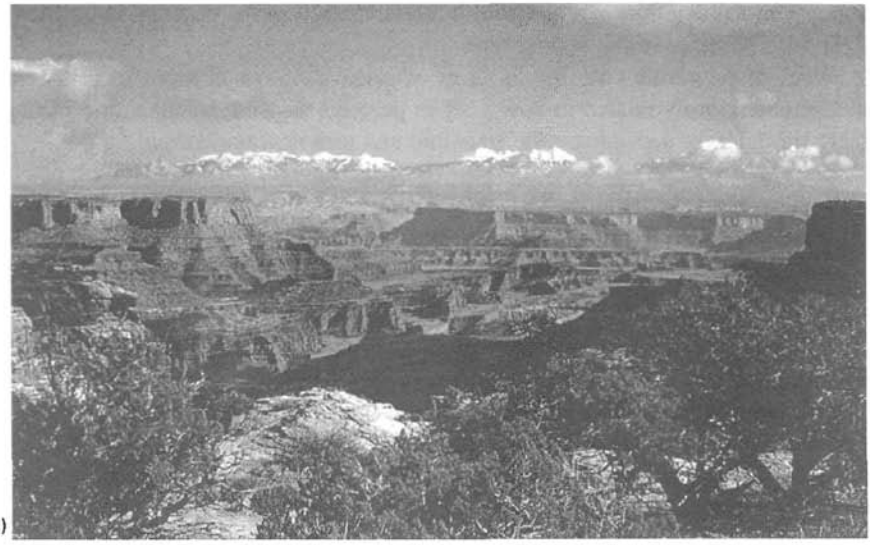


(a)

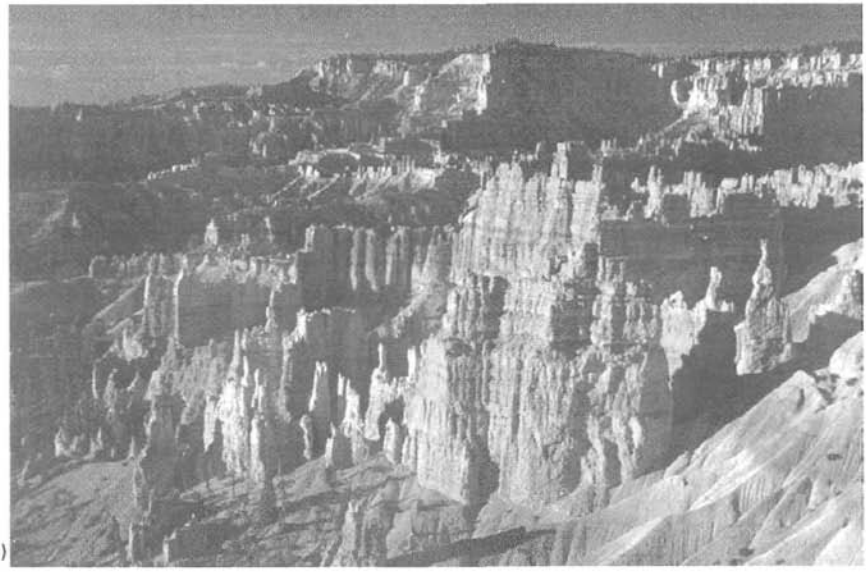


(b)

Figure 1 Photographs (*a*) through (*d*) show that, from a visual resource point of view, visibility is not how far a person can see but rather the ability of an observer to clearly see and appreciate the many and varied scenic elements in each vista. (*a*) The farthest scenic feature is the 130-km distant Navajo Mountain, as seen from Bryce Canyon National Park. (*b*) The La Sal Mountains, as seen from the Colorado River, are a dominant view from the distant horizon. (*c*) This view in Canyonlands National Park shows the highly textured foreground canyon walls against the backdrop of the La Sal Mountains. The La Sals are 50 km from the observation point. (*d*) Bryce Canyon as seen from Sunset Point. Notice the highly textured and brightly colored foreground features. See ftp site for color image.



(c)



(d)

Figure 1 Continued

and 1*d*, the observer may comment on the contrast detail of nearby geological structures or on shadows cast by overhead clouds.

Visibility, in the context of viewing scenic vistas, is more closely associated with conditions that allow appreciation of the inherent beauty of landscape features. It is important to be able to see and appreciate the form, contrast detail, and color of near and distant features. Therefore, visibility includes psychophysical processes and

concurrent value judgments of visual impacts, as well as the physical interaction of light with particles in the atmosphere.

Whether we define visibility in terms of visual range or in terms of some parameter more closely related to how visitors perceive a visual resource, the management of visibility depends on the scientific and technical understanding of:

- How aerosols are dispersed across land masses and into local canyons and valleys
- How they transform from a gas into particles that impair visibility
- How they interact with light
- The psychophysical processes involved in viewing scenic landscape features

Scientific understanding of some of these issues is more complete than others. The focus of this discussion is on developing a basic understanding of the interaction of light with aerosols and the psychophysical properties of the eye–brain system as they relate to visibility.

2 THEORY OF RADIATION TRANSFER AND VISIBILITY

The response of the human eye to radiant energy of different wavelengths is shown in Figure 2. The maximum response to a unit of energy is at 0.55 μm. When radiant energy is discussed in terms of the response of the human eye, photometric concepts and units are conventionally used. Conversely, when the entire radiation field of the sky is modeled or measured, radiometric units are employed. Usually, but not always, photometric parameters are derived from the more fundamental radiometric variables. Table 1 lists the various radiometric and corresponding photometric variables typically employed in radiation transfer calculations.

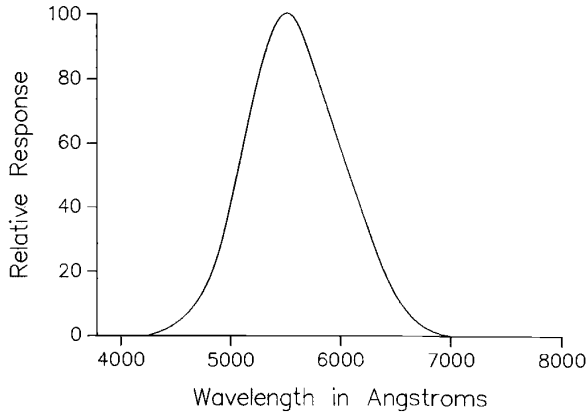


Figure 2 Spectral response of the human eye.

TABLE 1 Radiometric and Photometric Concepts and Units

Radiometric	Symbol	Units	Photometric	Symbol	Units
Radiant energy	U	joule	Luminous energy	Q	Talbot
Radiant flux	P	watt	Luminous flux	F	lumen
Radiant intensity	J	watt/steradian	Luminous intensity	I	lumen/steradian
Radiance	N	watt/m^2 steradian	Luminance	B	lumen/m^2 steradian
Irradiance	H	watt/m^2	Illuminance	E	lumen/m^2

Atmospheric Scattering and Extinction

The alteration of radiant energy as it passes through the atmosphere is due to scattering and absorption by gases and particles. The sum of scattering and absorption is referred to as the extinction coefficient. The effect of the atmosphere on the visual properties of distant objects theoretically can be determined if the concentration and characteristics of air molecules, particles, and absorbing gases are known throughout the atmosphere and most importantly along the line of sight between the observer and object. The extinction coefficient is made up of particle and gas scattering and absorption:

$$b_{\text{ext}} = b_{sg} + b_{ag} + b_{sp} + b_{ap} \tag{1}$$

where s, a, g, and p refer to scattering, absorption, gases, and particles, respectively.

Light scattering by gases is described by the Rayleigh scattering theory (vandeHulst, 1981). Important characteristics of Rayleigh scattering are:

- Its proportionality to molecular number density ($b_{sg} = 12\ \text{Mm}^{-1}$ at sea level and at 0.55 μm).
- The amount of scattered light varies as $1/\lambda^4$ where λ is the wavelength of light.
- Equal amounts of light are scattered in forward and backward directions.
- Light scattered at 90° is nearly completely polarized.

The only gas that is normally found in the atmosphere which absorbs light is nitrogen dioxide, NO_2. Absorption by NO_2 at 550 nm is $b_{ag} = 330[NO_2]$, where the units of b_{ag} are Mm^{-1} and the units of $[NO_2]$ are ppm (Nixon, 1940; Hodkinson, 1966). Furthermore, NO_2 absorbs more in the blue portion of the spectra than in the red portion. Therefore, NO_2 appears brown or yellowish if viewed against a background sky.

In most instances, particle scattering and absorption are primarily responsible for visibility reduction. Single-particle scattering and absorption properties can, with a number of limiting assumptions, be calculated using Mie theory (vandeHulst, 1981; Mie, 1908). However, before such calculations are carried out, appropriate boundary conditions must be specified. Typically aerosol models assume:

External Mixtures Particles exist in the atmosphere as pure chemical species that are mixed without interaction.

Multicomponent Aerosols Single particles are made up of two or more species.

Transfer of Radiant Energy

Visibility involves more than specifying how light is absorbed and scattered by the atmosphere. Important factors involved in seeing an object are outlined in Figure 3 and summarized below:

- Illumination of the overall scene by the sun, which includes illumination resulting from sunlight scattered by clouds and atmosphere as well as reflections by ground and vegetation
- Scene characteristics that include color, texture, form, and brightness
- Optical characteristics of intervening atmosphere:

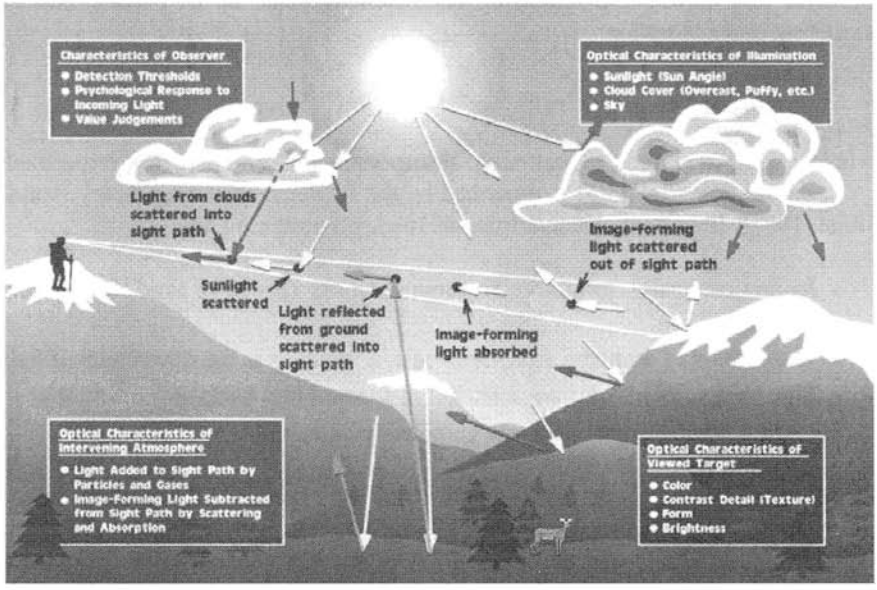


Figure 3 Important factors involved in seeing a scenic vista are outlined. Image-forming information from an object is reduced (scattered and absorbed) as it passes through the atmosphere to the human observer. Air light is also added to the sight path by scattering processes. Sunlight, light from clouds, and ground-reflected light all impinge on and scatter from particulates located in the sight path. Some of this scattered light remains in the sight path, and at times it can become so bright that the image essentially disappears. A final important factor in seeing and appreciating a scenic vista are the characteristics of the human observer. See ftp site for color image.

- Image-forming information (radiation) originating from landscape features is scattered and absorbed (attenuated) as it passes through the atmosphere toward the observer.
- Sunlight, ground-reflected light, and light reflected by other objects are scattered by the intervening atmosphere into the sight path.
- Psychophysical response of the eye–brain system to incoming radiation

Image-forming information is lost by the scattering of imaging radiant energy out of the sight path and absorption within the sight path, while ambient light scattered into the sight path adds radiant energy to the observed radiation field. This process is described by:

$$\underset{\text{(loss)}}{\frac{dN_r(\theta, \varphi, \mathbf{r})}{dr}} = -b_{\text{ext}} N_r(\theta, \varphi, \mathbf{r}) + \underset{\text{(gain)}}{N_*(\theta, \varphi, \mathbf{r})} \tag{2}$$

where $N_r(\theta, \varphi, \mathbf{r})$ is the apparent radiance at some vector distance, $\mathbf{r}$, from a landscape feature, $N_*(\theta, \varphi, \mathbf{r})$ (referred to as the path function) is the radiant energy gain within an incremental path segment, and $b_{\text{ext}} N_r(\theta, \varphi, \mathbf{r})$ is radiant energy lost within that same path segment. The atmospheric extinction coefficient (b_{ext}) is the sum of both atmospheric scattering (b_s) and absorption (b_a). Although not explicitly stated, it is assumed that each variable in, and each variable derived from, Eq. (2) is wavelength dependent. The parenthetical variables $(\theta, \varphi, \mathbf{r})$ indicate that N_r and N_* are dependent both on the direction of image transmission and on the position within the path segment. For the sake of brevity, the parenthetical variables will be dropped in following equations. When the postscript r is appended to any symbol, it denotes that the quantity pertains to a path of length r. The subscript 0 always refers to the hypothetical concept of any instrument located at zero distance from the object—as, for example, in denoting the inherent radiance of a surface. Prescripts identify the objects; the prescript b referring to background and l to landscape feature.

When N_r has some special value, N_q, such that $b_{\text{ext}} N_q = N_*$, then $dN_q/dr = 0$; N_q is independent of r and is commonly referred to as the equilibrium radiance. Therefore, for every path segment

$$\frac{dN_r}{dr} = -b_{\text{ext}}(N_r - N_q) \tag{3}$$

If N_q is constant, Eq. (3) can be integrated to yield

$$\frac{N_r - N_q}{N_0 - N_q} = T_r \tag{4}$$

where T_r is the transmittance over path length r and is given by

$$T_r = \exp - \int_0^r b_{\text{ext}} r' dr' \tag{5}$$

Rearranging Eq. (4) yields

$$N_r = N_0 T_r + N_q(1 - T_r) \tag{6}$$

where the first term on the right of Eq. (6) is the residual image-forming radiance, while the second term is the path radiance (airlight), N_r^*, which results from scattering processes throughout the sight path. The parameter N_∞^* is the sky radiance:

$$N_\infty^* = N_q(1 - T_\infty) \tag{7}$$

If T_∞ is approximately zero, then $N_q = N_\infty^* = N_s$ and

$$N_r^* = N_s(1 - T_r) \tag{8}$$

where N_s is sky radiance. Equation (8) allows for a simple approximation of N_r^* when N_s is known.

The explicit dependence of N_r^* on illumination and directional scattering properties of the atmosphere are best examined by considering

$$N_r^* = \int_0^r N_* T_r \, dr \tag{9}$$

where

$$N_* = h_s \sigma + \int_{4\pi} N \sigma \, d\Omega \tag{10}$$

The second term on the right-hand side is the contribution to N_* from sky, cloud, and earth radiance and $d\Omega$ is an element of solid angle. The parameter h_s is sun irradiance, and σ is the volume scattering function defined in such a way that

$$b_s = \int_{4\pi} \sigma \, d\Omega \tag{11}$$

Therefore, σ describes the amount of radiant energy (light) scattered in some direction, while the sum of radiant energy scattered in all directions is proportional to the scattering coefficient b_s. The amount of energy scattered out of and into a sight path over some incremental distance, Δr, is proportional to b_s. It is a fundamental optical property of the atmosphere. Its measurement and characterization have been the focus of a number of studies.

Contrast Transmittance in Real Space

Any landscape feature can be thought of as consisting of many small pieces, or elements, with a variety of physical characteristics. For instance, the reflectivity of an element as a function of wavelength, along with characteristics of the incident radiation, determines its color and brightness. The brightness of an element at some observing distance and at one wavelength is referred to as monochromatic apparent spectral radiance. The monochromatic apparent spectral radiance of any element is given according to Eq. (6) by

$$ {}_lN_r = T_r {}_lN_0 + N_r^* \tag{12} $$

where N_r^* is substituted explicitly for $N_q(1 - T_r)$. The subscript l indicates that the radiance is associated with a specific uniform landscape feature. In the early literature the subscript t (for target) was used instead of l because of the applicability of Eq. (12) and contrast to the seeing of military targets.

A scenic element is always seen against some background, such as the sky or another landscape feature. The apparent and inherent background radiance are related by an expression similar to Eq. (12)

$$ {}_bN_r = T_r {}_bN_0 + N_r^* \tag{13} $$

Subtracting Eq. (13) from Eq. (12) yields the relation

$$ ({}_lN_r - {}_bN_r) = T_r({}_lN_0 - {}_bN_0) \tag{14} $$

Thus, radiance differences are transmitted along any path with the same attenuation as that experienced by each image-forming ray.

The image-transmitting properties of the atmosphere can be separated from the optical properties of the object by the introduction of the contrast concept. The inherent spectral contrast, C_0, of a scenic element is, by definition,

$$ C_0 = ({}_lN_0 - {}_bN_0)/{}_bN_0 \tag{15} $$

The corresponding definition for apparent spectral contrast at some distance r is

$$ C_r = ({}_lN_r - {}_bN_r)/{}_bN_r \tag{16} $$

If Eq. (14) is divided by the apparent radiance of the background ${}_bN_r$ and combined with Eqs. (15) and (16), the result can be written as

$$ C_r = C_0 \frac{{}_bN_0}{{}_bN_r} T_r \tag{17} $$

Substituting Eq. (13) for ${}_bN_r$ and rearranging yields

$$\tau_r \equiv \frac{C_r}{C_0} = \frac{1}{1 + \dfrac{N_r^*}{{}_bN_0T_r}} \tag{18}$$

The right-hand member of Eq. (18) is an expression for the contrast transmittance, τ_r, of the path of sight. Equation (18) is the law of contrast reduction by the atmosphere expressed in the most general form. It should be emphasized that Eq. (18) is completely general and applies rigorously to any path of sight regardless of the extent to which the scattering and absorbing properties of the atmosphere or the distribution of lighting exhibit nonuniformities from point to point.

Visual Range Concept

Substituting Eq. (5) into Eq. (17) yields

$$C_r = C_0 \frac{{}_bN_0}{{}_bN_r} e^{-b_{\text{ext}}r}. \tag{19}$$

If an object is viewed against a background sky under uniform illumination conditions and through a uniform haze, ${}_bN_0/{}_bN_r = 1$ and Eq. (19) becomes

$$b_{\text{ext}} = -\frac{1}{r} \ln \frac{C_r}{C_0} \tag{20}$$

Equation (20) forms the basis for using teleradiometer contrast measurements for approximating the extinction coefficient. If C_0 and distance r are known, b_{ext} can be calculated.

The distance at which C_r approaches a threshold contrast of between -0.02 or -0.05 defines the visual range, V_r. If $|C_0| = 1$ (black object) and -0.02 is taken to be a threshold contrast, then Eq. (20) becomes

$$V_r = \frac{3.192}{b_{\text{ext}}} \tag{21}$$

Equation (21) allows visual range data to be interpreted in terms of extinction and vice versa, extinction measurements to be interpreted in terms of visual range. There is some debate as to what threshold contrast to use.

Equivalent Contrast

The above mathematical formalism is limited in that it does not account for human visual system response to edge sharpness between adjacent scenic features or to changes in contiguous contrast for features with varying size. More modern psycho-

physical perception threshold formalisms can be constructed to incorporate the eye–brain system response to variations in edge sharpness between landscape features as well as variation in spatial frequency of landscape scenic elements (Carlson and Cohen, 1978; Campbell and Robson, 1964; Campbell et al., 1968; Campbell and Kulikowski, 1986; Henry, 1977; Malm, 1985; Malm et al., 1987). Any approach that incorporates the human response to spatial frequencies (size and shape effects) is most easily handled using linear system theory. A first step is to develop a quantitative descriptor of the scene itself.

A scene can be decomposed into light and dark bars of various spatial frequencies and intensities whose brightness change is proportional to a sine wave function. Equivalent contrast, C_{eq}, is just the average contrast of those sine waves within specified frequencies. Therefore, equivalent contrast can be calculated either for all spatial frequencies or only for those frequencies to which the human visual system responds. Then C_{eq} can be used in human visual system models to estimate the probability that a human observer will notice a change in the appearance of a landscape feature as aerosols are added or removed from the atmosphere.

Contrast Transmittance in Spatial Frequency Space (Modulation Transfer Function)

In a derivation similar to the contrast transmittance derivation, it can be shown that the transmittance of equivalent contrast through the atmosphere in the presence of aerosols is given by

$$C_{eq,r} = C_{eq,0} M_{tf,a} \tag{22}$$

where

$$M_{tf,a} = \frac{1}{1 + \dfrac{N_r^*}{a_\infty T_r}} \tag{23}$$

and $C_{eq,r}$ and $C_{eq,0}$ are the equivalent contrast at distance r and 0, respectively, while $M_{tf,a}$ is the atmospheric modulation transfer function. The parameter a_∞, the average scene radiance, is the zero-order term in a two-dimensional Fourier decomposition of the scene radiance field.

Comparison of Eqs. (18) and (23) shows that if ${}_bN_0 = a_\infty$, then contrast transmittance in real and spatial frequency space is identical. In most cases, the feature within the image of interest is small compared with its surroundings, and average radiance, a_∞, is very nearly the same as background radiance ${}_bN_0$. This is a very satisfying result. Whether one is interested in using modern psychophysical spatial frequency models to examine how much aerosol can be introduced into the atmosphere before it is noticed or how image contrast is changed as a function of aerosol load, the calculation is reduced to understanding the dependence of the atmospheric

modulation transfer function, or contrast transmittance, on aerosol chemical and physical properties.

Dependence of Contrast Transmittance (τ_r) on Atmospheric Optical Variables

Because the contrast transmittance is the one variable that contains all the information required to describe how various physical descriptors of scenic landscape features are modified as a function of aerosol loading, illumination, and observer-vista geometry, it is of interest to examine how sensitive τ_r is to changes in atmospheric aerosol loading as a function of aerosol mass and average scene radiance. The average scene radiance, $\bar{N}$, was identified as a_∞ in Eq. (23).

Malm and Henry (1987) examined how the τ_r changes with changing image reflectivity, image distance, aerosol size distribution, and aerosol mass loading. For a sulfate aerosol, b_{ext} is almost entirely due to scattering and, as such, b_{ext} is proportional to aerosol mass. Therefore, the variation of τ_r with respect to b_{ext} is proportional to its variation with respect to aerosol mass. Figures 4*a* and 4*b* show $s \equiv |\Delta\tau_r \Delta b_{\text{ext}}|$ as a function of b_{ext}.

Figure 4*a* corresponds to a typical sulfate aerosol mass size distribution, scattering angle $\theta_s = 15°$, and $N_0 = 0.13N_s$, where N_s is the Rayleigh sky radiance. Figure 4*b* is also for a sulfate aerosol but with $\theta_s - 125°$ and $N_0 = 0.5N_s$. An immediately evident trend shown in Figures 4*a* and 4*b* is that there is a distance where S is maximum; S decreases to zero as $R \to 0$ and as $R \to \infty$. Secondly, the distance at which S is maximum increases as N_0 increases (brighter landscapes). In a forward scattering situation where landscapes are in a shadow ($C_0 \approx -0.90$), S is maximum in the 5- to 10-km range. Although not explicitly shown in Figure 4*a*, in a backscatter geometry ($\theta_s = 125°$), the most sensitive distance is still around 5 to 10 km if the landscape is dark. However, the maximum sensitivity drops by about a factor of 2 and is not nearly as sensitive to distance. On the other hand, Figure 4*b* shows that when the landscape is highly reflective and illuminated ($C_0 \approx 0.50$ and $\theta_s = 125°$) the distance of maximum sensitivity increases, is quite sensitive to background b_{ext}, and remains sensitive to changes in b_{ext} long after dark targets have lost their sensitivity (dark targets will have disappeared, while bright targets can still be seen).

Figure 5 examines in more detail the relative contribution of N_r^* and T to S. Figure 5 shows contributions of $N_r^*$and T to S for the case shown in Figure 4*a* at $T = 10$ km (forward scattering, sulfate aerosol, and dark target). Changes in N_r^* are primarily responsible for changes in $M_{\text{tf},a}$ as aerosol is added or subtracted from a clean atmosphere. As background aerosol loading is increased (larger b_{ext}), the relative importance of T to S increases to a point where T dominates the effect on S. However, it should be emphasized that this only occurs after the $M_{\text{tf},a}$ has increased to a point where landscape features would be barely visible. Figure 5*b* shows N_r^* and T contributions to S for the Figure 4*b* case at $R = 70$ km (backscatter, sulfate aerosol, and bright target). With this geometry, attenuation of image-forming information, T, is responsible for much of the change in $M_{\text{tf},a}$. In fact, N_r^* can

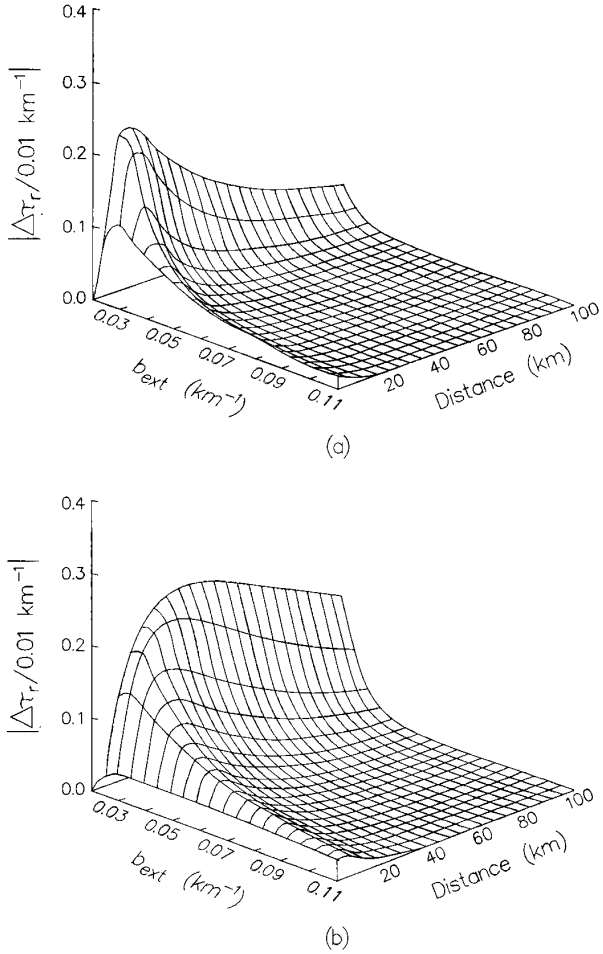


Figure 4 The sensitivity of the absolute value of contrast transmittance ($\Delta\tau_r/\Delta b_{ext}$) plotted as a function of extinction coefficient and distance to landscape feature. (*a*) Scattering angle $\theta_s = 15°$, shadowed vista $\bar{N}_0 = 0.13\ N_s$, and sulfate aerosol, and (*b*) scattering angle $\theta_s = 125°$, illuminated vista $\bar{N}_0 = 0.5\ N_s$, and sulfate aerosol.

decrease as b_{ext} increases and compensate slightly (contribute to cause $M_{tf,a}$ to increase) for decreases in T.

The foregoing discussion shows that the effect of increasing b_{ext} (aerosol concentration) for a scattering aerosol in almost all situations causes $M_{tf,a}$ to decrease. However, under forward scattering situations where targets tend to be dark, N_r^*

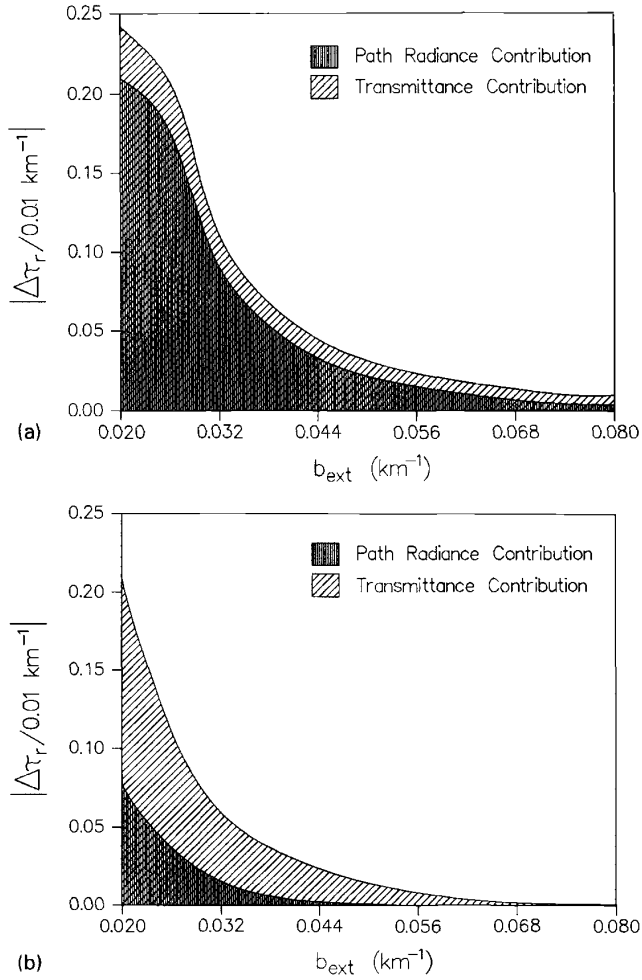


Figure 5 Sensitivity (S) expressed as changes in the modulation transfer function increase of $b_{ext} = 0.01\ km^{-1}$ plotted against b_{ext} for a sulfate aerosol. In (a) $\theta_s = 15°$, $R = 70$ km, and $\bar{N}_0 = 0.13\ N_s$, while in (b) $\theta_s = 125°$, $R = 70$ km, and $\bar{N}_0 = 0.5\ N_s$. Parts (a) and (b) show the relative contributions of path radiance and atmospheric transmittance to changes in $M_{tf,a}$ as a function of b_{ext}.

dominates changes in $M_{tf,a}$. On the other hand, when looking at brightly colored landscape features with the sun behind the observer's back (backscatter), the relative importance of N_r^* to visibility becomes smaller and changes in N_r as a result of increased b_{ext} are more dependent on image-forming radiance being attenuated over

the sight path. However, for a specific scene under static illumination conditions, contributions of N_r^* and T to change in $M_{tf,a}$ as a function of aerosol concentration tend to track each other.

Because most research to date has focused on apportionment of b_{ext}, and therefore T, to aerosol species, it is fortunate that for scattering aerosols, such as sulfates, an understanding of this relationship yields significant insight into how aerosols affect visibility under a wide range of viewing conditions. However, under not uncommon circumstances, the major cause of visibility degradation can be associated with path radiance, and path radiance explicitly requires knowledge of the volume scattering function in addition to b_{ext}. Almost no effort has been expended on examining how path radiance is affected as a function of aerosol characteristics or on apportioning path radiance to aerosol species. Aerosols that absorb light contribute to path radiance differently than aerosols that only scatter light (such as sulfates), so the impact of scatterers and absorbers on path radiance is not additive. Conversely, the effect of scattering and absorbers in b_{ext} is additive. Therefore, when appreciable concentrations of light-absorbing particles or gases are present, knowledge of just b_{ext} (transmittance) may not be adequate to describe changes in visibility.

The concepts discussed above are summarized in Figure 6. Those variables enclosed in the box on the left side of Figure 6 are dependent on illumination observer geometry, while those on the right are not. Path radiance, a geometry-

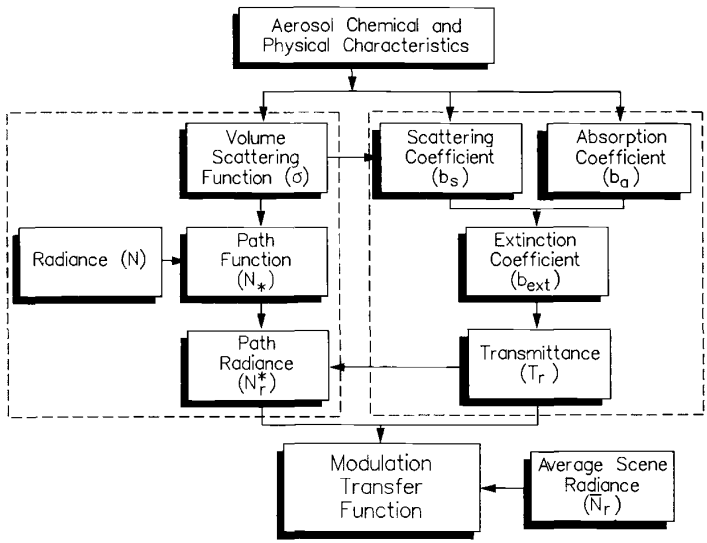


Figure 6 Flow diagram showing how aerosol physical-chemical characteristics relate to the optical variables required to completely specify the atmospheric modulation transfer function.

dependent variable, is combined with atmospheric transmittance, a geometry-independent parameter, and average scene luminance to yield contrast transmittance or modulation transfer function.

3 VISIBILITY IMPAIRMENT

Aerosols introduced into the atmosphere can result in visibility impairment that is manifested in two distinct ways: first, as a general alteration in the appearance of landscape features such as color, contiguous contrast between adjacent geologic features, etc., and secondly the aerosol haze may become visible in and of itself. Haze may be visible by the contrast or color difference between itself and its background, or (at great enough optical depths) uniform haze manifests itself as a semitransparent curtain that can be seen or perceived as a separate hazy entity disassociated from landscape features. Henry (1987) has referred to the phenomenon as atmospheric transparency, which is psychophysical in nature, and different from atmospheric transmittance.

Perceptibility Parameters for Quantification of Layered Haze (Plume Blight)

Figure 7 illustrates two situations in which a layered haze is visible: (a) when viewed against the sky and (b) when viewed against terrain features. In both cases, the layered haze will be visible as a distinct, horizontal layer if it is sufficiently brighter or darker than the viewing background.

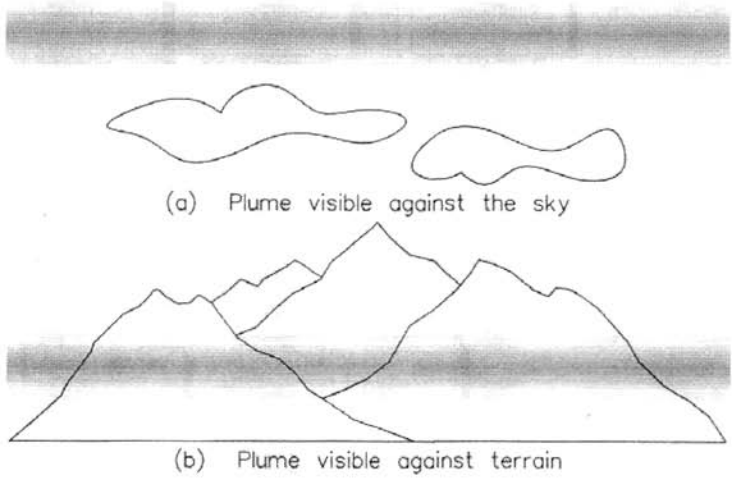


Figure 7 Two viewing situations in which plumes may be visible.

The simplest way to characterize the relative brightness (or darkness) of plumes is through the use of plume contrast:

$$C = \frac{{}_pN_r - {}_bN_r}{{}_bN_r} \tag{24}$$

where ${}_pN_r$ and ${}_bN_r$ are the spectral radiances of the plume and its background, at some distance, r, and at wavelengths in the visible spectrum ($0.4 < \lambda < 0.7$ μm). A plume is visually perceptible only if it creates a nonzero contrast at different wavelengths in the visible spectrum greater than an observer's perceptibility threshold (generally in the range of ± 0.01 to ± 0.05).

An object can be perceived because it has a brightness different from that of the background or because it has a different color. Gases and particles in the atmosphere can give rise to coloration by their light-scattering properties (blue sky or white clouds) or by altering the color of objects seen through them (brown coloration due to NO_2). Several schemes have been used to quantify color. The Commission Internationale de l'Eclairage (CIE) has set colorimeter standards that form the basis of the CIE system of color specification. The most popular CIE index is the so-called ΔE parameter that not only quantifies differences in color but also differences in brightness. However, the CIE method, while accurate and acceptable for a laboratory situation, may not adequately represent color differences in a natural setting. In any case, a ΔE of 1 is a just noticeable difference in color and/or brightness in a laboratory setting and ΔE of 4 can be easily seen by the casual observer.

Layered Haze Thresholds. Psychophysical research (Cornsweet, 1970; Faugeras, 1979; Hall and Hall, 1977; Henry, 1986; Howell and Hess, 1978; Malm et al., 1987; Ross et al., 1987) has documented the fact that the human eye–brain system is most sensitive to spatial frequencies of approximately three cycles/degree (cpd). Spatial frequency is defined as the reciprocal of the distance between sine-wave crests (or troughs) measured in degrees of angular subtense of a sine-wave grating. Thus, spatial frequency has units of cycles/degree. Any pattern of light intensities, whether it is a sine wave, square wave, step function, or any other pattern, can be resolved by Fourier analysis into a sum of sine-wave curves of different magnitude and frequency. For instance, a rough estimate of the primary spatial frequency of a Gaussian plume can be made as follows. If it were assumed that a Gaussian distribution is nearly identical to a sine-wave pattern, a 2° width of the plume would correspond to the period of the sine wave. The spatial frequency would be the inverse of this, or 0.5 cpd. Figure 8 illustrates several estimates of the sensitivity of the human visual system to sine/square-wave gratings and single Gaussian and square-wave stimuli with various spatial frequencies.

The sensitivity of the human eye–brain system drops off significantly at high spatial frequency (due to visual acuity) and also to a lesser extent at low spatial frequency (i.e., broad, diffuse objects). The human visual system is more sensitive to images with sharp, distinct edges (e.g., square waves) than to images with diffuse, indistinct edges (e.g., sine waves or Gaussian plumes).

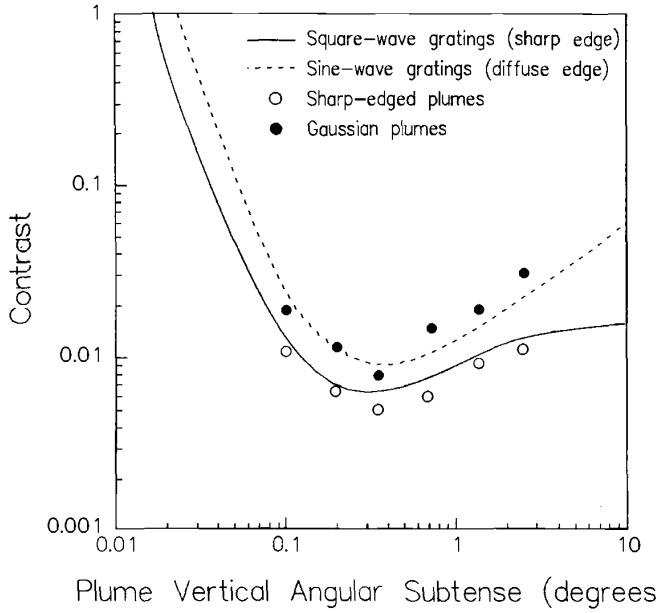


Figure 8 Sensitivity curves as reported by Howell and Hess (1978) for sine- and square-wave ratings and for sharp-edged (Malm et al., 1987) and Gaussian plumes (Ross et al., 1990).

Ross et al. (1997), based on an extensive literature review, designed a laboratory study to develop the information necessary to predict the probability of detection of plumes with a known size, shape, and contrast. The strategy taken was to develop probability of detection curves for computer-generated plume stimuli that encompasses the various plume geometries that could be encountered in the "real" world and interpolate between these measured thresholds to develop estimates for plumes with other shapes and geometries. In each case, the protocol for observer detection was the same for all experiments, the surround was kept at the same brightness, edge effects were dealt with uniformly, and stimuli representative of Gaussian plume brightness profiles were used.

Sixteen subjects were used for a full-length plume experiment. The stimuli used consisted of plumes with vertical angular sizes of 0.09°, 0.18°, 0.36°, 0.72°, 1.44°, and 2.88° and a horizontal angular extent of 16°. Contrast values of 0.050, 0.040, 0.030, 0.020, 0.017, 0.015, 0.013, 0.011, and 0.005 were used for all sizes. Figure 9 shows the predicted probability of detection curves. As plume contrast increases, the probability of detecting the plume increases. If a plume has a modulation contrast of greater than about 0.01 it will be detected nearly 100% of the time for all size plumes. Furthermore, these curves show that the size of the plume is quite impor-

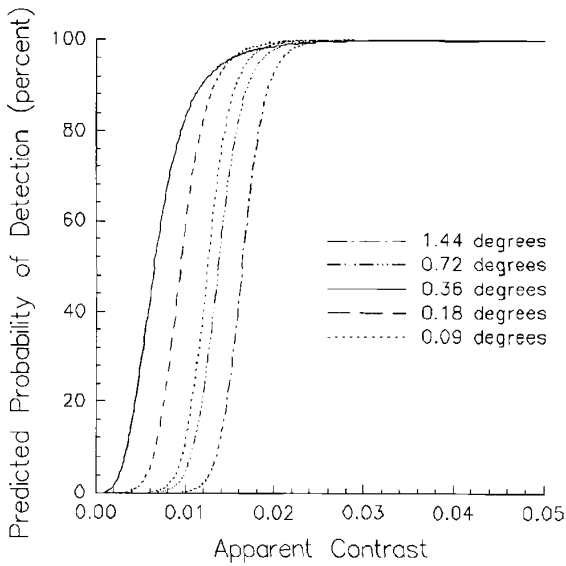


Figure 9 Predicted probability of detection curves for one subject used in the full-length plume study.

tant! Plumes that subtend an angle of about 3° can be detected more easily than plumes that are larger or smaller. Results for circular and oval-type plumes with Gaussian edges were similar but required higher contrasts to be detected.

To more clearly see how the three shapes compared, the modulation contrast corresponding to 50% probability of detection for each shape is plotted against plume size in Figure 10. Notice that the general trend for all stimuli is the same, with plumes subtending about a 3° width being the easiest to detect. However, observers are most sensitive to full-length plumes and least sensitive to circular stimuli with the oval plumes being intermediate. The full length, oval, and circular plume contrast threshold data have been incorporated into a linear interpolation algorithm that allows plumes of any size to be estimated.

Other studies have been carried out for brighter than background-layered hazes. They identified a 70% detection threshold contrast of 0.02 using photographs of a natural scene with light-colored layered hazes, which varied in size. The evidence for ΔE thresholds is not as clear-cut. The data of Jaeckel (1973) and Malm et al. (1980) support 70% detection thresholds for ΔE of three, while the estimates of Latimer et al. (1978) and the more recent data of Malm et al. (1987) and Henry and Matamala (1990) suggest a ΔE threshold of less than 1. This work is summarized in Table 2.

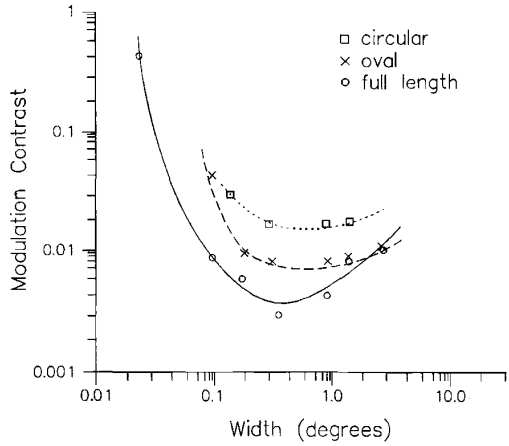


Figure 10 Threshold modulation contrast is plotted as a function of plume width in degrees for full-length, oval, and circular plumes. The human observer is most sensitive to all plumes if they have a width that is about 3°. Plumes larger or smaller than about 3° require increased contrast to be seen.

Perceptibility Parameters for Quantification of Uniform Haze Impairment

Whereas work discussed in the previous sections has emphasized detection thresholds of layered hazes, specifically plumes, other researchers have concentrated their efforts in establishing the change in image appearance required to just notice a difference in image sharpness.

Early work focused on establishing the just noticeable difference between a scene where an object viewed against the same background could just be seen and one where that object could not be identified. This threshold work was carried out in the context of establishing the "threshold" contrast for visual range determination.

More recent work has been directed toward incorporating results of basic psychophysical measurements into models that will predict the change in display modulation transfer function (MTF) required to evoke a just noticeable difference (JND) in display image sharpness. Displays of interest were television-type video displays. One model, the quadratic detection model (QDM), relies on the calculation of the image mean square luminance fluctuation, termed the image modulation depth. Henry (1979) and Henry et al. (1981) have suggested that modulation depth may be appropriate visibility indices because they incorporate all of the information content contained in a scenic vista.

Malm and Pitchford (1989) have suggested using the concept of a just noticeable change (JNC) in the appearance of a landscape feature as a psychophysical variable that relates directly to human perception. A JNC corresponds to the amount of absorbing gas or atmospheric particulate matter required to evoke a noticeable

change in the appearance of a particular landscape. The effect of a change in aerosol concentration can then be expressed as the number of JNCs between landscape appearance under current conditions versus the appearance after a change in emissions. Malm and Pitchford (1989) have suggested using the QDM to predict a JNC; however, any psychophysical model relating changes in aerosol concentration to human eye–brain visual thresholds could be used for this purpose. It is emphasized that none of the currently used psychophysical models have been field validated.

More recently, Pitchford and Malm (1994) have proposed the deciview scale, which is based on the fact that all detection threshold models and experiments show that above contrasts of about 0.02, a just noticeable change in contrast, is directly proportional to the initial contrast, $\Delta C = LC$, where L is a proportionality constant. By assuming the availability of sensitive scenic targets at every distance, it can then be demonstrated that any specific fractional change in extinction coefficient is equally perceptible regardless of baseline visibility conditions. The index is defined so that its scale, which is expressed in deciview (dv), is linear with respect to fractional changes in extinction and is given by, $\mathrm{dv} = 10 \ln(b_{ex}/0.01\ \mathrm{km}^{-1})$, where extinction is expressed in inverse kilometers. A 1-dv change is about a 10% change in extinction.

Application of the Quadratic Detection Model. Typical scenes are made up of features that are quite varied with respect to size, shape, and luminance level. However, some attempts have been made to classify scenic structure into broad categories such as form, line, and texture. Form refers to large shapes seen either against sky or other uniform background, while line is usually associated with appearance of rivers or similar geological features. Texture refers to the periodic contrast associated with sparsely populated trees seen against a uniform background, varied geologic features, or other similar higher frequency scenic structures.

Studies investigating eye fixation and eye motion as observers look at pictures show that pictorial areas with little modulation receive very little attention, while higher modulated scenic features receive more (Boswell, 1975). Since high-contrast edges are most sensitive to changes in atmospheric modulation transfer function and since the discrimination of an atmospheric modulation change in a frequency-specific channel is a minimum when the contrast in that channel is largest, it can be concluded that high-contrast edges are good patterns for predicting the relationship between just noticeable changes in scenic appearance and increases in atmospheric aerosol load.

For many typical scenes, a JNC is equivalent to a change in atmospheric modulation of approximately 0.06. Figure 11 shows a typical JNC surface for an 80% reduction in atmospheric extinction as a function of observer distance and atmospheric extinction, assuming a change in MTF of 0.06 is perceptible. The scattering angle is 15°, $a_\infty = N_s$ where N_s is sky brightness, and the initial contrast, C_0, is equal to -1.0. A typical aerosol mass size distribution with typical chemical properties was assumed.

There are some general features that show up in all JNC surfaces. For any given distance there is a background extinction that is most sensitive to an incremental

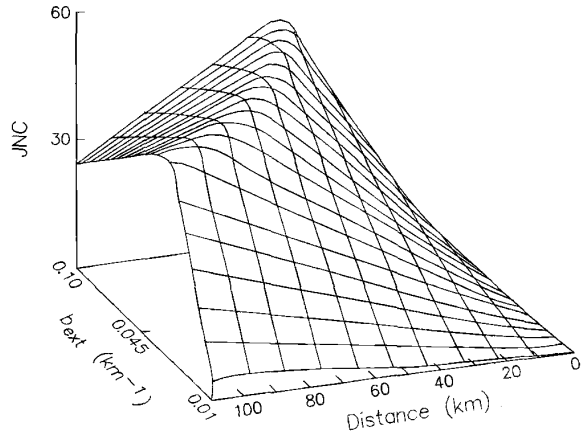


Figure 11 Just noticeable change surface plotted as a function of observer distance and atmospheric background extinction. The surface corresponds to a reduction of background extinction of 80%.

change in extinction, and for any given extinction there is observer distance that is most sensitive to extinction change. Secondly, for any given observer distance, the sensitivity of a scene to incremental reductions in atmospheric extinction drastically reduces as background extinction increases; and finally, the distance where the scene is most sensitive to a change in extinction decreases as background extinction increases.

Human Judgments of Visual Air Quality

The previous section discussed methodologies for establishing the change in atmospheric particulate loading required to be noticeable either as a layered haze or as a change in scenic quality. It should be emphasized that calculations of detection thresholds and JNCs are statements about changes in information content in an image. JNC changes in the appearance of an image are not necessarily good indicators of judged image quality. For instance, a change in 10 JNCs in a scene with low overall contrast may not be judged to have the same change in image quality as 10 JNCs in a high-contrast scene.

Studies by Malm et al. (1980, 1981), Latimer et al. (1980, 1983), Middleton et al. (1984), Stewart et al. (1983), and Hill (1990) have established relationships between judgments of image quality of natural scenes and various atmospheric and vista parameters such as mountain/sky contrast, solar angle, extinction coefficient, sky color, and percent cloud cover. Latimer et al. (1980) had observers judge scenic beauty (SBE) and visual air quality (VAQ) for a number of eastern and western national park vistas as they appeared under a variety of illumination and meteorological conditions. The results of their study were mixed and in some cases contra-

dictory. In Latimer et al. (1980, p. 113), they conclude that "to different extents for different vistas, ratings of VAQ and SBE both increase with increasing visual range." In Latimer et al. (1983, pp. 49–50), they conclude that "ratings of SBE of a given vista were independent of visual range unless there was a dominant distant landscape feature in the landscape scenery." Since the visual range calculation "normalizes" out specific unique characteristics of vistas, these results are not surprising. The Latimer studies did conclude that changes in illumination did have a considerable effect on SBE ratings. Middleton et al. (1984) also concluded that illumination was important to VAQ judgments and were able to show at one site that there is a good correlation between VAQ and $\ln(b_{scat})$, where b_{scat} is the atmospheric scattering coefficient. Additionally, Hill (1990) emphasizes that color is extremely important to judgments of scenic beauty.

Malm et al. (1980) examined the relationship between VAQ and vista contrast. They showed that, under fixed illumination and meteorological conditions, apparent vista contrast of the most distant vista element was a good prediction of VAQ judgments. The study also showed that changes in foreground color (due to change in illumination), addition of clouds, or snow cover caused the VAQ ratings to be higher but did not cause the sensitivity of VAQ to change in vista contrast change. Malm et al. (1980) also presented a model of human perception of VAQ. The model is based on the observation that ratings of VAQ are proportional to the sum of the fraction of each scenic element subtended by various landscape features multiplied by the atmospheric transmittance between that landscape feature and observer. It was shown that when a single landscape feature, void of color and textural detail, dominates the perceived change in visual air quality, the model predicts a linear relationship between VAQ and the apparent contrast of that landscape feature (contrast of form).

Several researchers have found that judgments of photographs can be used as surrogates for judgments made in the field provided the experiments have been properly designed. This is an important finding since one way to reduce the per-observation cost of obtaining judgment-based measurements of visual air quality is to use judgments of photographs rather than field observations. For example, Stewart et al. (1984) found that although visual air quality tends to be judged slightly worse in photographs than in the field, the relative differences among scenes are approximately the same whether visual air quality is judged from photographs or in the field.

The implication of the visual air quality perception research described in the preceding paragraphs is that there are a number of variables such as sun angle, cloud cover, and scene composition that are firmly integrated into judgments of aesthetic value of a scenic resource. Therefore, studies designed to assess social, psychological, or economical value associated with a given change in atmospheric particulate concentration must be designed in such a way that these confounding variables do not affect the outcome of the experiment. For instance, a number of experiments have been carried out using photographs of landscape features under a variety of air quality conditions as the stimulus. To avoid extraneous variables such as sun angle from affecting the study, it is essential that the study be carried out

using photographs taken at the same time of day and under similar lighting conditions.

4 EXAMPLES OF VISIBILITY IMPAIRMENT

The camera can be an effective tool in capturing the visual impact that pollutants have on a visual resource. Figures 12*a* to 12*d* show the effect that various levels of uniform haze have on a Glacier National Park vista. These photographs were taken of the Garden Wall from across Lake McDonald. Figures 13 and 14 show similar hazes of vistas at Mesa Verde and Bryce Canyon National Parks. The Chuska Mountains in Figure 13 are 95 km away. Navajo Mountain in Figure 14 is 130 km distant. This photograph should be compared with Figure 1*a*, a photograph of Navajo Mountain taken on a day in which the particulate concentration in the atmosphere was near zero.

Under stagnant air mass conditions aerosols can be "trapped" and produce a visibility condition usually referred to as layered haze. Figure 15 shows Navajo Mountain viewed from Bryce Canyon National Park with a bright layer of haze that extends from the ground to about halfway up the mountain. Figure 16 is a similar example of layered haze but with the top portion of the mountain obscured.

Figure 12 Effect of regional or uniform haze on a Glacier National Park vista. The view is of the Garden Wall from across Lake McDonald. Atmospheric particulate concentrations associated with photographs (*a*), (*b*), (*c*), and (*d*) correspond to 7.6, 12.0, 21.7, and 65.3 $\mu g/m^3$. See ftp site for color image.

(b)

(c)

Figure 12 (*Continued*)

Figure 17 is a classic example of plume blight. In plume blight instances, specific sources such as those shown in Figure 18 emit pollutants into a stable atmosphere. The pollutants are then transported in some direction with little or no vertical mixing.

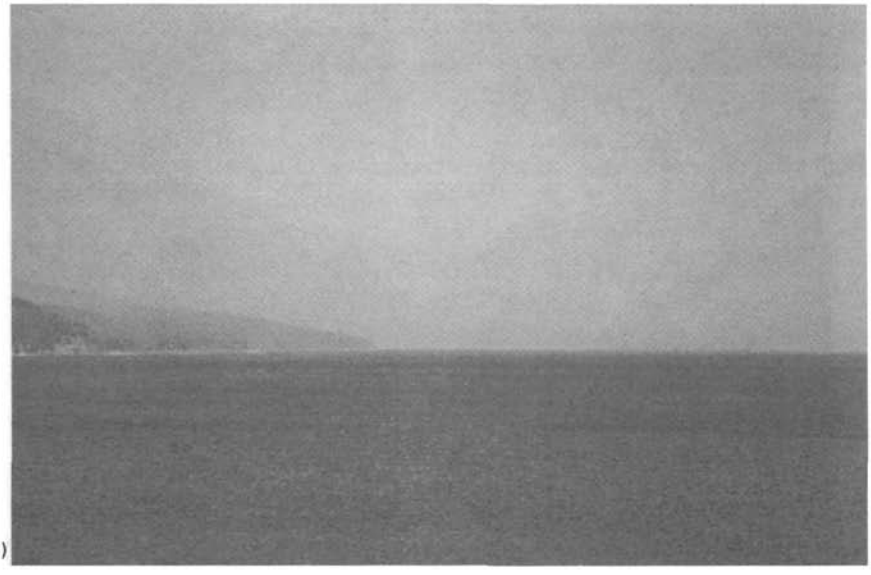


Figure 12 (*Continued*)

Figure 13 Effects of uniform haze on the Chuska Mountains as seen from Mesa Verde National Park. The atmospheric particulate concentration on the day this photograph was taken corresponded to 1 $\mu g/m^3$. See ftp site for color image.

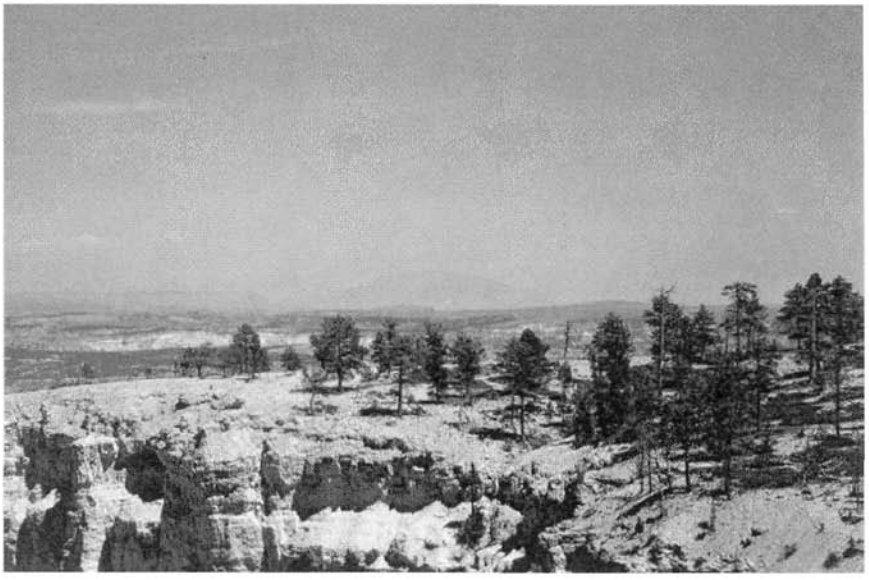


Figure 14 Uniform haze degrades visual air quality at Bryce Canyon National Park. The 130-km distant landscape feature is Navajo Mountain. Atmospheric particulate concentration on the day this photograph was taken is 3 $\mu g/m^3$. See ftp site for color image.

Figure 15 Navajo Mountain as seen from Bryce Canyon, showing the appearance of layered haze. The pollutants are trapped in a stable air mass that extends from the ground to about half-way up the mountain side. See ftp site for color image.

Figure 16 Photograph of Navajo Mountain similar to Figure 15 but with a suspended haze layer that obscures the top portion of the mountain. See ftp site for color image.

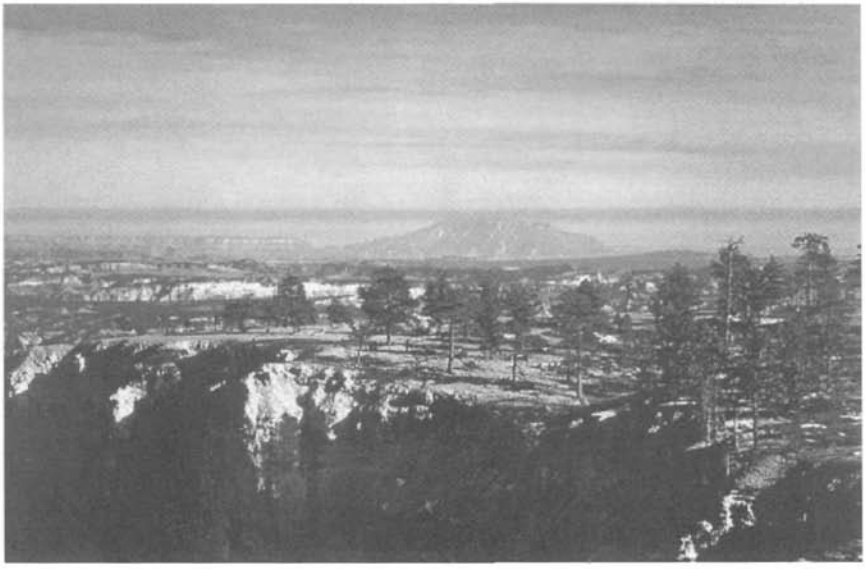


Figure 17 Classic example of "plume blight." The thin, dark plume on Navajo Mountain results from a point source emitting particulate matter into a stable atmosphere. See ftp site for color image.

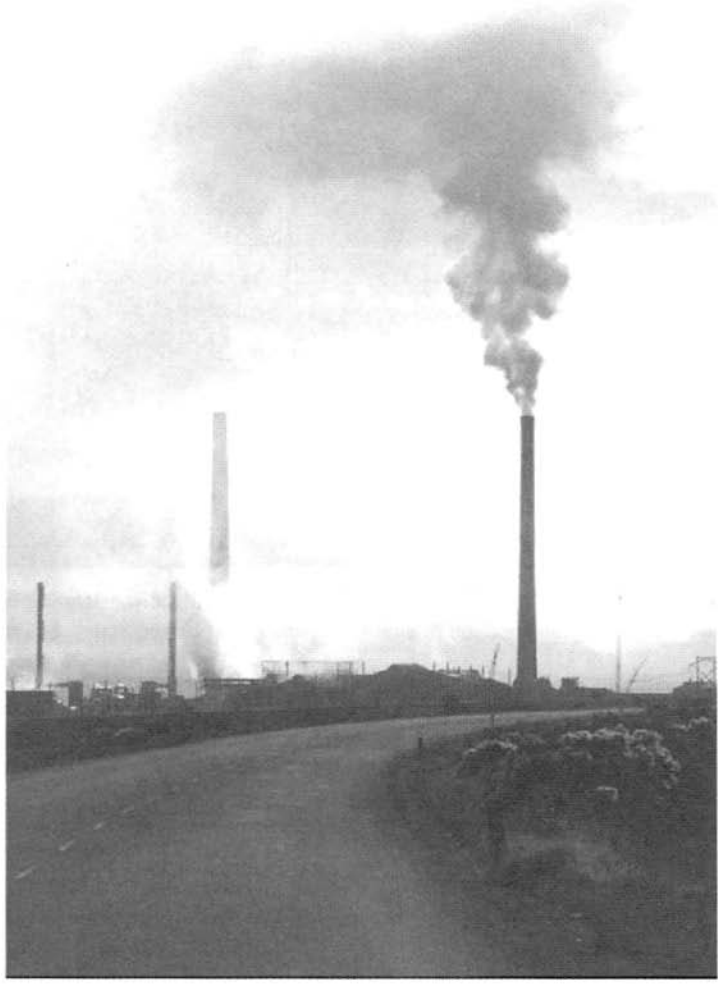


Figure 18 Example of one kind of point source that emits pollutants into the atmosphere. See ftp site for color image.

Figures 19, 20, 21, and 22 show other layered haze conditions that frequently occur at Grand Canyon and Mesa Verde National Parks. At Mesa Verde (Figure 22), much of the pollution comes from urban areas and the Four Corners and San Juan Power Plants, while at the Grand Canyon layered hazes are associated with smoke and nearby coal-fired power plants.

Figures 23 and 24 show the appearance of plumes containing carbon. In both of these cases the pollutants are being emitted from forest fires. However, Figure 23 shows the appearance of a specific forest fire plume, while Figure 24 shows the effect of viewing a vista through a concentration of particles containing carbon. In this instance, the vista is the north wall of the Grand Canyon as seen from the top of

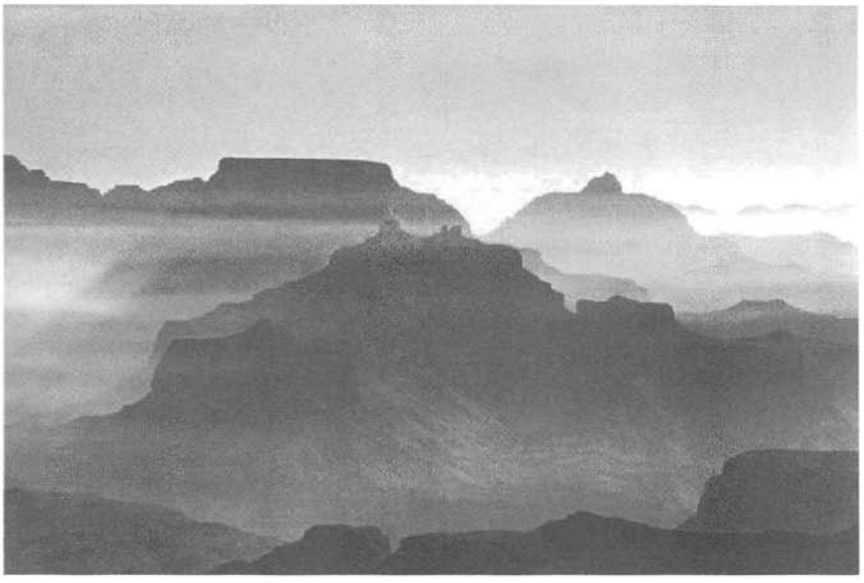


Figure 19 Smoke trapped by an inversion layer in the Grand Canyon. During the winter months, inversions are quite common in almost all parts of the United States. See ftp site for color image.

San Francisco Peaks in northern Arizona. Notice the overall "graying" and reduction of contrast of the distant scenic features. Remember that carbon absorbs all wavelengths of light and scatters very little. Thus the scene will always tend to be darkened.

Figure 20 Example of power plant emissions trapped in an air inversion layer in the Grand Canyon. See ftp site for color image.

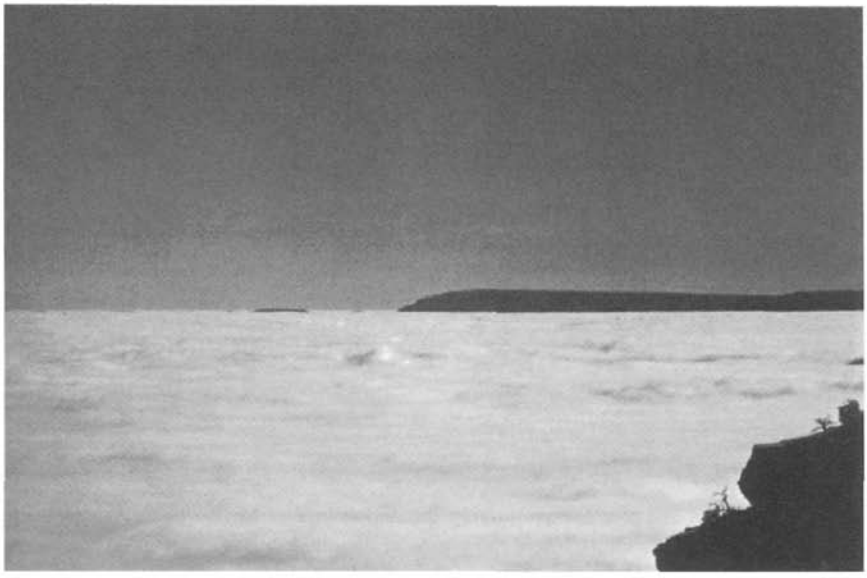


Figure 21 Effects of inversion layer in Grand Canyon. In this case, a cloud has formed within the canyon walls. See ftp site for color image.

Figure 22 Effects of layered haze trapped in front of the Chuska Mountains as viewed from Mesa Verde National Park. This condition occurs 30 to 40% of the time during winter months. See ftp site for color image.

Figure 23 Forest fire plume exemplifying the appearance of carbon particles and demonstrating the effect of lighting. Where the plume is illuminated, it appears gray, but identical particles in the shadow of the plume appear dark or almost black. See ftp site for color image.

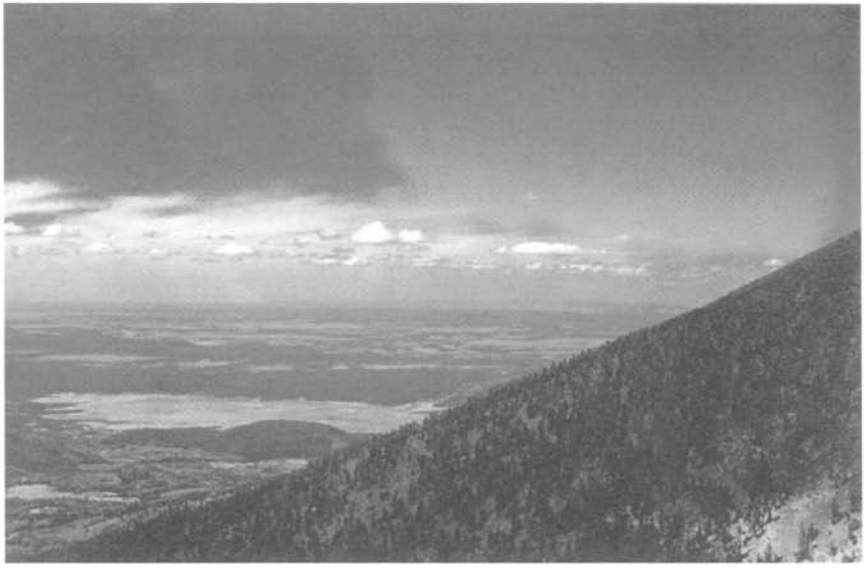


Figure 24 Example of how light-absorbing particles (in this case carbon) affect the ability to see a vista. Carbon absorbs all wavelengths of light and generally causes a "graying" of the overall scene. Shown here is the north wall of the Grand Canyon as seen from the top of the San Francisco Peaks in northern Arizona. See ftp site for color image.

Figure 25 shows the effects of illumination on the appearance of power plant plumes. The two plumes on the left are particulate plumes, while the two plumes on the right consist of water droplets. The plume on the far right, which is illuminated by direct sunlight, appears to be white. The second, identical water droplet plume, which is shaded, appears dark. The amount of illumination can have a significant effect on how particulate concentrations appear.

Figure 26 demonstrates how the effect of nitrogen dioxide gas (NO_2), in combination with varied background illumination, can combine to yield a very brown atmospheric discoloration. If a volume of atmosphere containing NO_2 is shaded and if light passes through this shaded portion of the atmosphere, the light reaching the eye will be deficient in photons in the blue part of the spectrum. As a consequence, the light will appear brown or reddish in color. However, if light is allowed to shine on, but not through, that same portion of the atmosphere, scattered light reaches the observer's eye and the light can appear to be gray in nature. Both of these conditions are shown in Figure 26. On the right side of the photo clouds shade the mixture of NO_2 and particulates. The same atmosphere, illuminated because the cloud cover is not present, appears almost gray in the middle portion of the photograph.

Effects of illumination are further illustrated in Figures 27*a*, and 27*b*. Figure 27 is an easterly view of the La Sal Mountains in southeastern Utah as seen from an elevated point that is some 100 km distant. The photograph in Figure 27*a* was taken at 9:00 A.M., while the photograph shown in Figure 27*b* was taken later in the day.

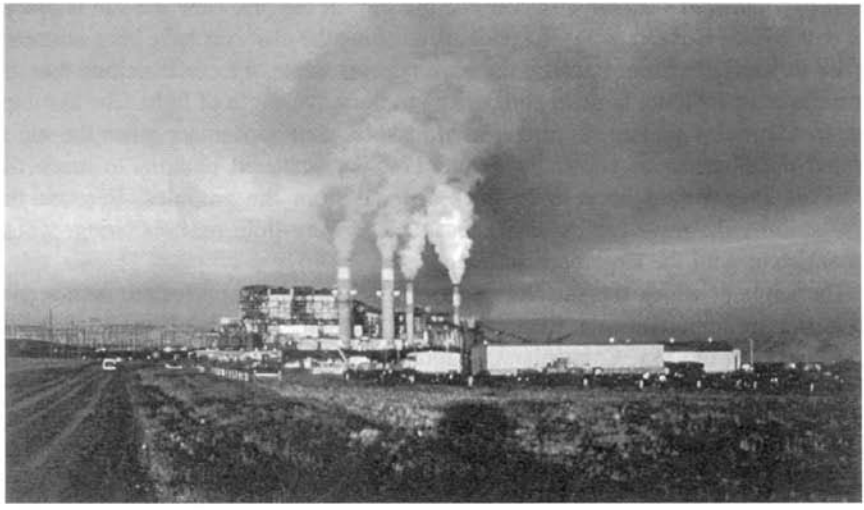


Figure 25 The effect of illumination on the appearance of plumes. The two plumes on the right are identical in terms of their chemical makeup, in that they are primarily water droplets. However, the far right plume is directly illuminated by the sun and the plume second from the right is shaded. The first plume appears white and the second appears almost black. The two plumes on the left are fly-ash plumes. See ftp site for color image.

Figure 26 (see color insert) The brown discoloration resulting from an atmosphere containing nitrogen dioxide (NO_2) being shaded by clouds but viewed against a clear blue sky. Light scattered by particulate matter in the atmosphere can cominate light absorbed by NO_2, causing a gray or blue appearing haze (left side of photograph). See ftp site for color image.

These photographs show how these views, or vistas, appear when obscured by a layer of haze. In the first view the haze layer appears white, but the same air mass viewed later in the day has a dark gray appearance. This effect is entirely due to the geometry involved with the observer and the sun. In the first view the sun is low in the eastern sky. Consequently, the photons reaching the observer have been scattered in the forward direction. Because the haze appears white, we can conclude that the particles must be quite large in comparison to the wavelength of light. The assumption that particles are large is further reinforced by their appearance when the sun is behind the observer as shown in Figure 27*b*. For scattered photons to reach the observer, they would have to be backscattered from the particles. Because the haze appears dark, we can conclude that there is very little backscattering, which is consistent with the large particle hypothesis.

The angle at which the sun illuminates a vista or landscape feature (sun angle) plays another important role. Figures 28*a* to 28*d* exemplify this effect. The view is from Island in the Sky, Canyonlands National Park, looking out over Canyonlands with its many colorful features toward the 50 km distant La Sal Mountains. Figure 28*a* shows how the canyon appears when it is in total shadow (6:00 A.M.). Figures 28*a* to 28*c* show a progressively higher sun angle until in Figure 28*d* the scene is entirely illuminated. In each case, the air quality is the same. The only change is in the angle with which the sun illuminated the vista. There are primarily two reasons for the apparent change in visual air quality. First, at higher sun angles, there is less scattering of light by the intervening atmosphere in the direction of the observer. Second, the vista reflects more light; consequently, more image-forming information (reflected photons from the vista) reaches the eye. The contrast detail and scene are enhanced.

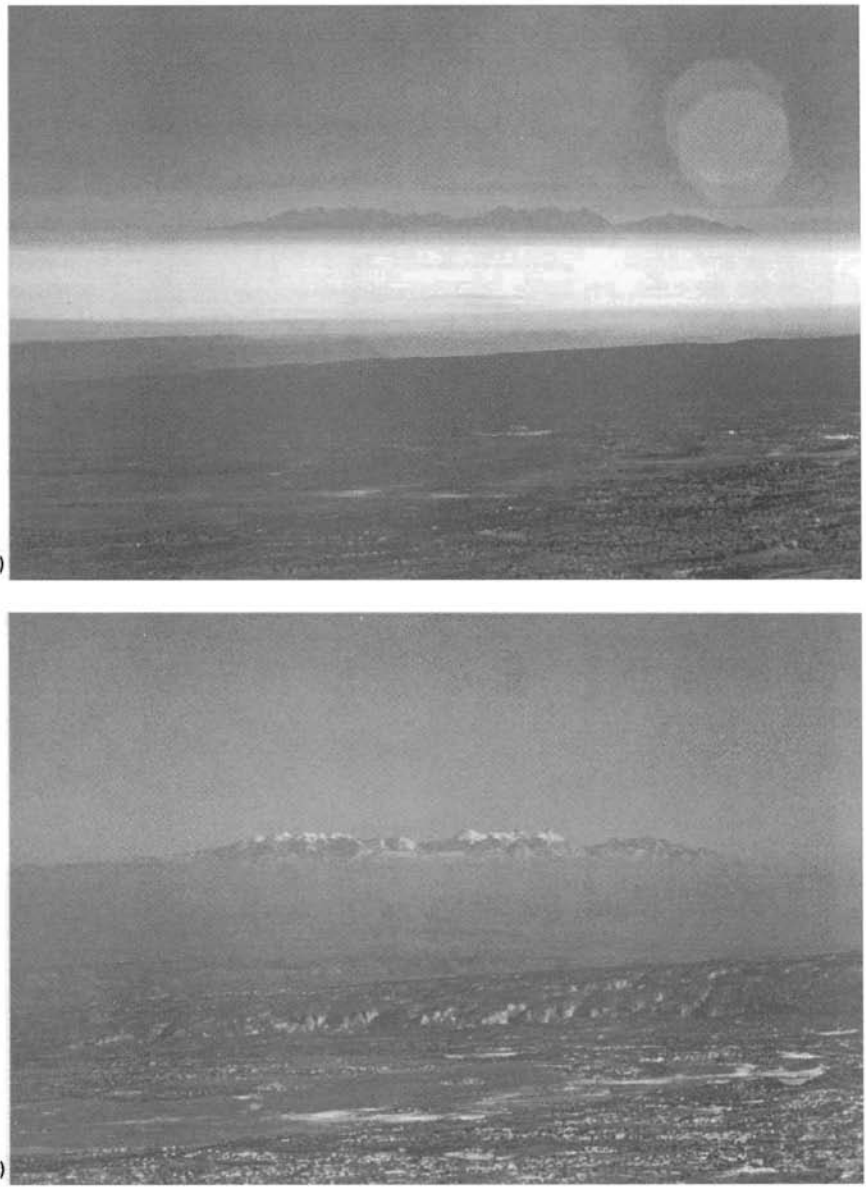


Figure 27 Photographs show how the same haze trapped in an inversion layer looks under forward and backscatter conditions. In (*a*) under forward scattering conditions (morning), the haze appears white; in (*b*) the identical haze, viewed in the afternoon during backscatter conditions, is dark or gray. Because most of the light energy is scattered in the forward direction (white haze), it can be concluded that the particles must be quite large in comparison to the wavelength of light. See ftp site for color image.

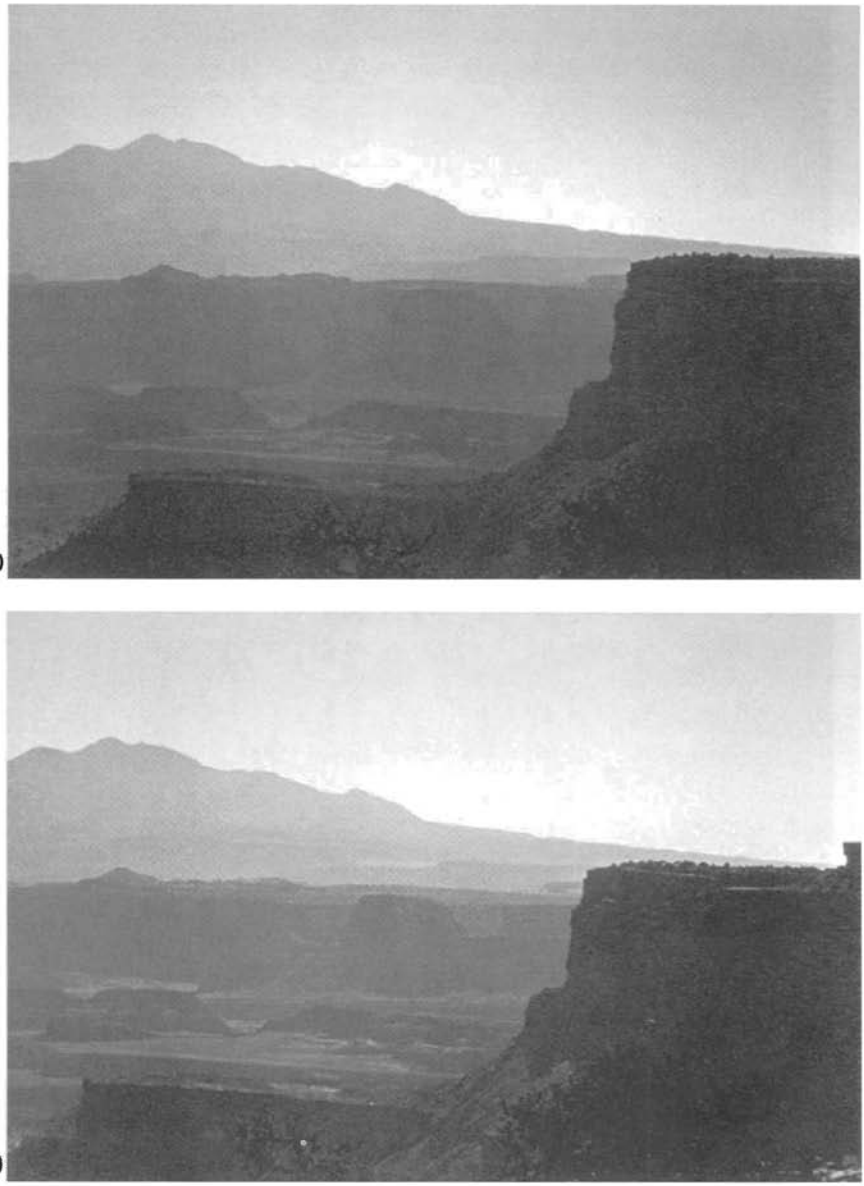
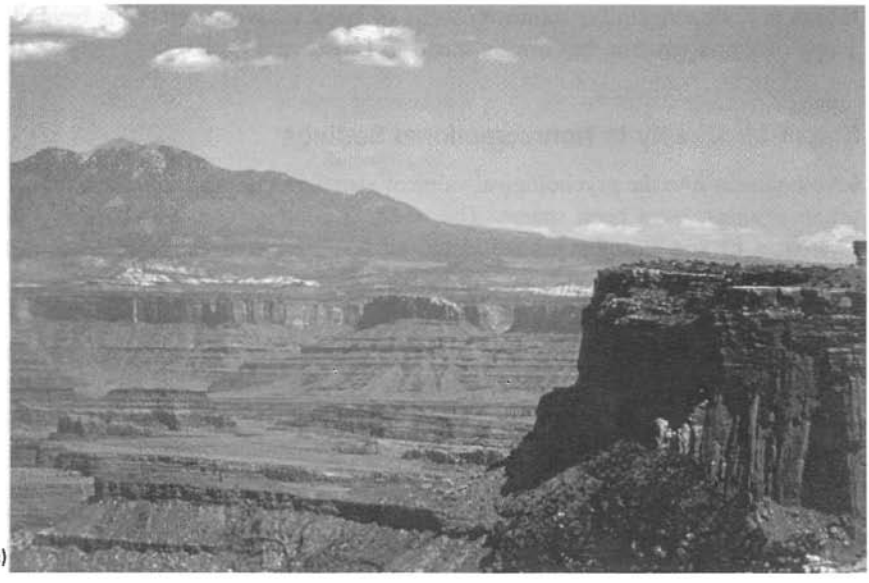
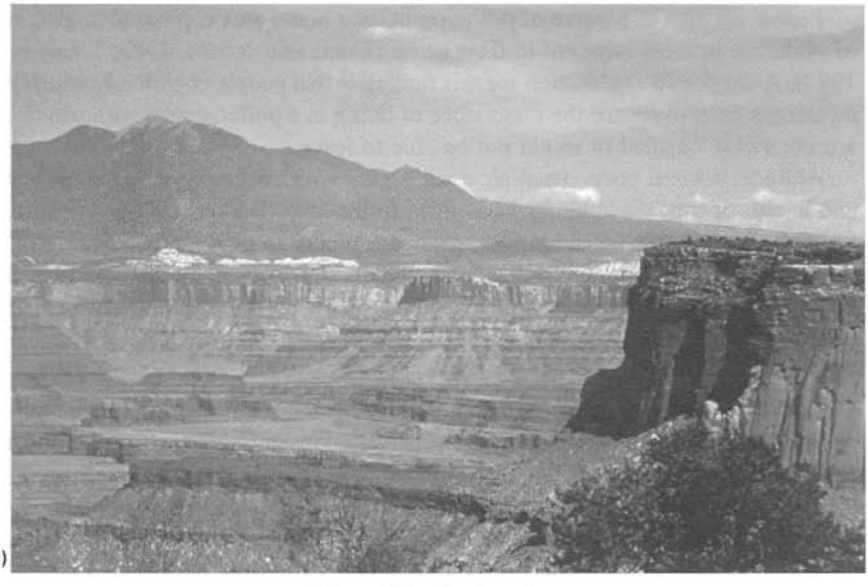


Figure 28 Four photographs showing the effect of shifting sun angle on the appearance of a vista as seen from Island in the Sky, Canyonlands National Park. In each photograph, the air quality is the same. In (*a*) (6:00 A.M.) the sun angle–observer–vista geometry results in a large amount of scattered air light (forward scattering) added to the sight path, but minimal amount of imaging light reflected from the vista. Figures 28*b* and 28*c* show a progressively higher sun angle until in Figure 28*d*, the scene is entirely illuminated. Scattered light is minimized and reflected; imaging light is at a maximum. See ftp site for color image.



(c)



(d)

Figure 28 Continued

5 VALUE OF GOOD VISUAL AIR QUALITY

Efforts to define and quantify the value of good visual air quality have generally followed two courses. One emphasis has been on monetary costs to resource degradation and human health. The other emphasis has been on the psychological value of visual air quality in the context of recreational and nonrecreational settings. A

thorough review regarding monetary value of good visual air quality is beyond the scope of this discussion but can be found in Brown and Callaway (1990).

Visual Air Quality in Nonrecreational Settings

Investigations into the psychological value of visual air quality in nonrecreational, or urban settings, have been sparse. The research conducted in this area examined awareness of and attitudes toward visual air quality and investigated relationships between visual air quality, stress, and human behavior.

Public Perception of Visual Air Quality. Survey research of public awareness of visual air quality using direct questioning typically reveals that 80% or more of the respondents are aware of poor visual air quality, and that poor visibility and media publicity are the primary factors that precipitate the awareness (Cohen et al., 1986). These surveys have also shown that awareness is not uniform across the general population of a given area. Persons with higher income and educational levels tend to be more aware of poor visual air quality than those with lower income and educational levels.

People are also less aware of pollution in their home area compared to awareness of pollution in areas adjacent to their home (Evans and Jacobs, 1982; Evans et al., 1982). A suggested explanation for this finding is that people cognitively adjust their awareness level to reduce the dissonance of living in a polluted area with which they are otherwise satisfied or might not be able to leave.

Attitudes toward poor visual air quality vary with socioeconomic status, health, and length of time an individual has lived in the area (Barker, 1976). Affluent and well-educated people consider poor visual air quality to be a more serious problem than others. People who are not economically tied to sources of air pollution, have respiratory ailments, or are new to an area also show the strongest negative reactions to reduced VAQ.

Visual Air Quality and Stress. Reduced visual air quality is an ambient environmental stressor because it is a relatively constant and unchanging situation over which one has little direct control (Campbell, 1983). The associated stress and lack of control is chronic, not salient, and may be manifested in heightened levels of anxiety, tension, anger, fatigue, depression, and feelings of helplessness (Evans et al., 1987; Ziedner and Shechter, 1988). How one deals with this stress is dependent on coping behavior and ability to adapt. The relationship between stress due to poor visual air quality and mental health is poorly understood. However, results from a study conducted by Rotton and Frey (1982) showed that as visual air quality decreased, emergency calls for psychiatric disturbances increased.

Visual Air Quality and Behavior. Evans et al. (1982) found that persons who recently moved to Los Angeles from areas with good visual air quality consistently reduced outdoor activities during periods of reduced visual air quality compared

with longer-term residents. Studies have also reported reduced altruism and increased hostility and aggression during periods of poor air quality (Cunningham, 1979; Jones and Bogat, 1978; Rotton et al., 1979). The relationship between aggression, hostility, and visual air quality is curvilinear with feelings of aggression and hostility increasing to a certain point and then dropping off and yielding to a desire to withdraw and escape from the situation. Evans and Cohen (1987) suggest that individuals adjust to poor visual air quality through adaptation and coping behaviors by altering their judgment of air quality based on current and previous exposure.

Visual Air Quality in Recreational Settings

During the past decade, an experience-based demand model has been developed to assess demand for recreational opportunities. The model incorporates visitor demand for activities, for social/physical/managerial site attributes, and for the realization of specific psychological satisfactions.

The model was used to investigate the psychological value of good visual air quality at Grand Canyon, Mesa Verde, Great Smoky Mountains, Mount Rainier, and Everglades National Parks using on-site interviews and mail-back surveys. The purpose was to evaluate the importance of visual air quality relative to other park attributes, to determine if visitors were accurately aware of changes in visibility, and to ascertain whether relationships existed between visual air quality and visitor satisfaction (Ross et al., 1985, 1987).

Importance of Good Visual Air Quality. The importance of good visual air quality to park visitors was evaluated by having visitors rate how important specific park attributes were to their recreational experience. Cluster analysis was used to statistically identify similar types of attributes based on response patterns. Grand Canyon National Park's attributes, their corresponding mean importance scores, and cluster formation are shown in Figure 29. The "clean, clear air" attribute ranked third in importance and combined with attributes that are descriptive of a clean, natural setting, which, as a group, were slightly more important than the cluster of view-related attributes. This indicates that visitors interpret "clean, clear air" as being an integral part of the cleanliness of the park and as such, an important part of the overall recreational experience sought at Grand Canyon.

The importance of a natural, clean environment with clean air was not unique to Grand Canyon visitors. Figure 30 shows that similar findings resulted from the other studies regardless of park location or overall theme. The cleanliness attribute cluster, which included "clean, clear air," was the most important cluster at all five parks.

Visitor Awareness of Visual Air Quality. In a random sample, nearly 1800 visitors at Grand Canyon National Park were asked during an interview if they were aware of any haze and, if so, how hazy they thought it was. Results from correlation analysis between awareness of haze and standard visual range measures showed that visitors' awareness of haze increased as visibility decreased. Correlation coefficients

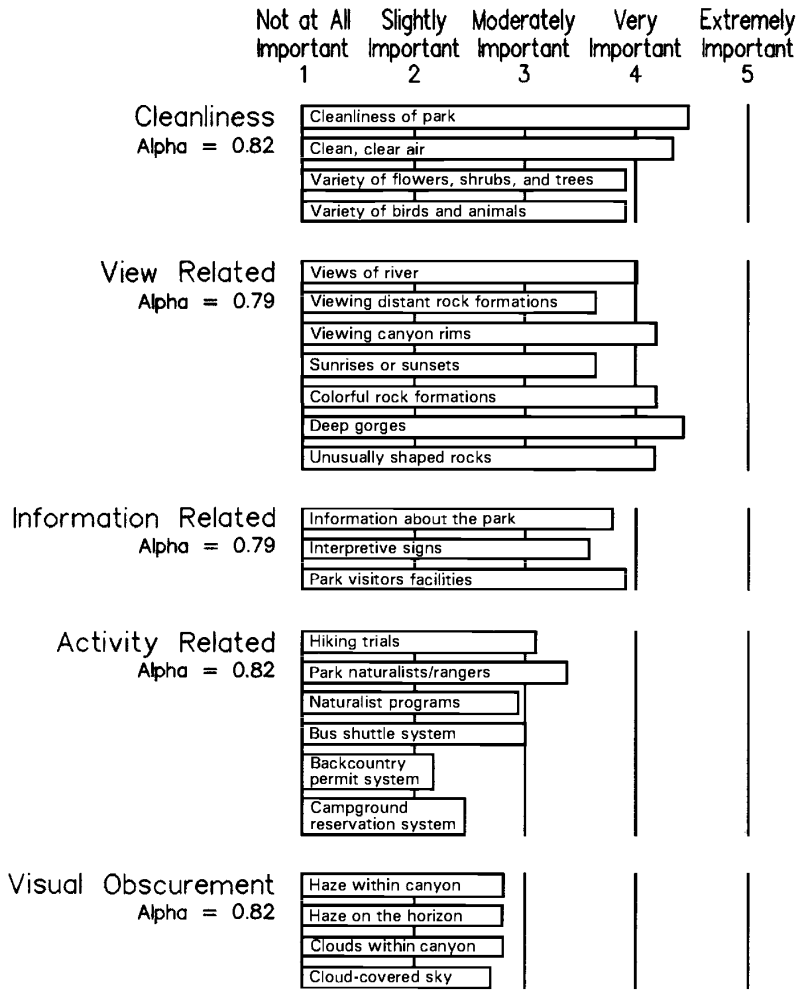


Figure 29 Attributes, attribute mean scores, attribute clusters, and attribute cluster mean scores for the Grand Canyon National Park visitor survey. See ftp site for color image.

were also calculated between visitor awareness of haze and ratings on enjoyment of the view, impact of haze on overall park enjoyment, and satisfaction with the "clean, clear air" attribute. Results showed that as awareness of reduced visibility increased, enjoyment with the view, overall park enjoyment, and satisfaction with the "clean, clear air" attribute decreased.

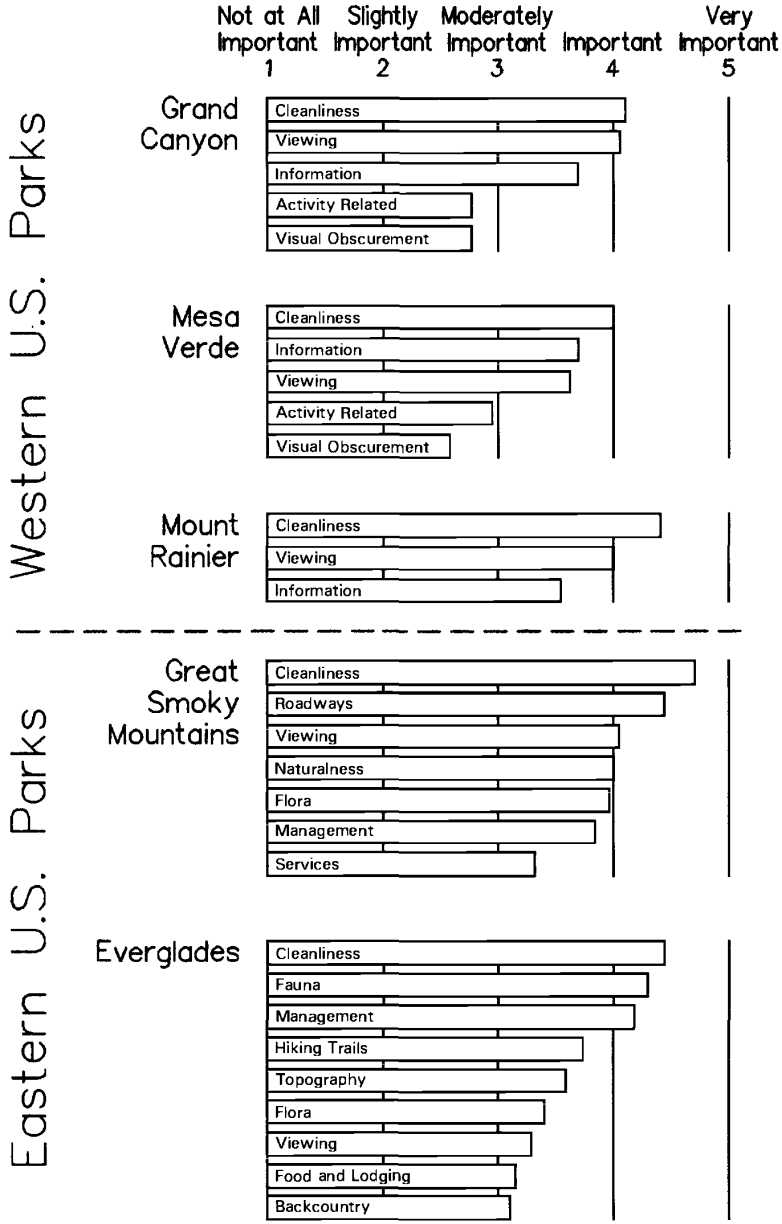


Figure 30 Relative importance of attribute clusters at five national parks. See ftp site for color image.

TABLE 2 Summary of Contrast and Color Change Threshold Data

Contrast	ΔE	Percent Detection	Edge	Reference
0.003[a]	—	50	Sharp	Blackwell (1946)
0.014	—	?	Sharp	Lowry (1931, 1951)
0.007[b]	—	?	Sharp	Howell and Hess (1978)
0.009[b]	—	?	Diffuse	
0.016[c]	—	?	Sharp	
—	1	30	Sharp	Jaeckel (1973)
—	2	50	Sharp	
—	3	70	Sharp	
—	4	90	Sharp	
0.006	1	10	Diffuse	Malm et al. (1980)
0.009	1.5	25	Diffuse	
0.014	2.3	50	Diffuse	
0.02	3.3	75	Diffuse	
0.025	4.2	90	Diffuse	
0.01	—	90	Sharp	Loomis et al. (1985)
0.005[d]	—	70	Sharp	Malm et al. (1985)
0.010[e]	—	70	Sharp	
0.020[f]	—	70	Diffuse	Ross et al. (1988)
0.007[d]	—	70	Diffuse	Ross et al. (1990)
0.025[e]	—	70	Diffuse	

[a]The most sensitive contrast reported for largest size of stimulus and largest luminance and longest response time evaluated (probably the minimum possible threshold).
[b]The most sensitive contrast reported at a spatial frequency of 3 cycles/degree.
[c]Threshold contrast for sharp objects at low spatial frequencies.
[d]Minimum threshold for 0.36° wide plumes.
[e]Maximum threshold for all size plumes tested.
[f]Threshold contrast reported for light-colored, diffuse edge hazes of varying size.

Visual Air Quality and Recreational Behavior. A laboratory study conducted by Malm et al. (1984) at Grand Canyon National Park examined how visual air quality might affect visitor behavior. Participants examined sets of photographs with different levels of visual air quality and indicated how they would be willing to spend a given amount of time either driving to a lookout point or touring an archaeological site. The study concluded that subjects place a high value on visual air quality and would be willing to significantly alter behavior for increased visual air quality. For example, subjects would be willing to spend an additional 2.5 h driving time to view a dominant distant landscape for a 0.01 km^{-1} reduction in atmospheric extinction. The study also showed that vistas that lacked color and texture were insensitive to increases in atmospheric extinction.

Disclaimer. The assumptions, findings, conclusions, judgments, and views presented herein are those of the author and should not be interpreted as necessarily representing official National Park Service policies.

REFERENCES

Barker, M. Planning for environmental indices: Observer appraisals of air quality, in K. Craig and E. Zube (Eds.), *Perceiving Environmental Quality: Research and Applications*, 175–203, Plenum, New York, 1976.

Blackwell, H. R., Contrast thresholds of the human eye, *J. Opt. Soc. Am.*, *36*, 624, 1946.

Boswell, G. T., *How People Look at Pictures*, University of Chicago Press, Chicago, IL, 1975.

Brown, G. M., and J. M. Callaway, Methods for valuing acidic deposition and air pollution efforts, National Acid Precipitation Assessment Program (NAPAP): State of Science, Report No. 27, Washington, DC, 1990.

Campbell, F. W., Ambient stressors, *Environ. Behavior*, *15*, 355–380, 1983.

Campbell, F. W., and J. J. Kulikowski, Orientational selectivity of the visual cell of the cat, *J. Physiol. (London)*, *187*, 437, 1986.

Campbell, F. W., B. Cleveland, G. F. Cooper, and C. Enroth-Cogell, The spatial selectivity of the visual cells of the cat, *J. Physiol. (London)*, *198*, 237, 1968.

Campbell, F. W., and J. G. Robson, Application of Fourier analysis to the modulation response of the eye, *J. Opt. Soc. Am.*, *54*, 581A, 1964.

Carlson, C. R., and R. W. Cohen, *Image Descriptors for Displays: Visibility of Displayed Information*, RCA Laboratories, Princeton, NJ, 1978.

Cohen, S., G. W. Evans, D. Stokols, and D. S. Krantz, *Behavior, Health, and Environmental Stress*, Academic, New York, 1986.

Cornsweet, T., *Visual Perception*, Academic, New York, 1970.

Cunningham, M., Weather, mood, and helping behavior: Quasi-experiments with the sunshine Samaritan, *J. Per. Soc. Psych.*, *37*, 1947–1956, 1979.

Evans, G. W., and S. Cohen, Environmental stress, in D. Stokols and I. Altman (Eds.), *Handbook of Environmental Psychology*, 571–602, Wiley, New York, 1987.

Evans, G. W., and S. V. Jacobs, Air pollution and human behavior, in G. W. Evans (Ed.), *Environmental Stress*, 105–132, Cambridge University Press, New York, 1982.

Evans, G. W., S. V. Jacobs, and N. B. Frager, Behavioral responses to air pollution, in A. Baum and J. Singer (Eds.), *Advances in Environmental Psychology*, Vol. 4, 237–270, Erlbaum, New York, 1982.

Evans, G. W., S. V. Jacobs, D. Dooley, and R. Catalano, The interaction of stressful life events and chronic strains on community mental health, *Am. J. Comm. Psych.*, *15*, 23–24, 1987.

Faugeras, O. D., Digital color image processing within the framework of a human visual model, *IEEE Trans. Acoust. Speech Sig. Process*, *ASSP-27*, 380–393, 1979.

Hall, C. F. and E. L. Hall, A nonlinear model for the spatial characteristics of the human visual system, *IEEE Trans. Syst. Man. Cybernet. SMC-7*, 161–170, 1977.

Henry, R. C., The application of the linear system theory of visual acuity to visibility reduction by aerosols, *Atmos. Environ.*, *11*, 697, 1977.

Henry, R. C., The Human Observer and Visibility—Modern psychophysics applied to visibility degradation, in *View on Visibility: Regulatory and Scientific*, 27–35, Air Pollution Control Association, Pittsburgh, PA, 1979.

Henry, R. C., Improved predictions of plume perception with a human visual system model, *J. Pollut. Control Assoc.*, *36*, 1353–1356, 1986.

Henry, R. C., Psychophysics, visibility, and perceived atmospheric transparency, *Atmos. Environ.*, *21*, 159–164, 1987.

Henry, R. C., and L. V. Matamala, Prediction of color matches and color differences in the outdoor environment, in C. V. Mathai (Ed.), *Transactions of Visibility and Fine Particles*, 554–561, Air and Waste Management Association, Pittsburgh, PA, 1990.

Henry, R. C., J. F. Collins, and D. Hadley, Potential for quantitative analysis of uncontrolled routine photographic slides, *Atmos. Environ.*, *15*, 1959, 1981.

Hill, A. C., Measuring How Landscape Color Affect Aesthetic Value, in *Transactions of Visibility and Fine Particles*, Air and Waste Management Association, Pittsburgh, PA, 570–581, 1990.

Hodkinson, J. R., Calculations of color and visibility in urban atmospheres polluted by gaseous NO_2, *J. Air Water Pollut. Int.*, *10*, 137–144, 1966.

Howell, E. R., and R. F. Hess, The functional area for summation to threshold for sinusoidal gratings, *Vision Res.*, *18*, 369–374, 1978.

Jaeckel, S. M., Utility of color-difference formulas for match acceptability decisions, *Appl. Opt.*, *12*, 1299–1316, 1973.

Jones, J. W., and G. A. Bogat, Air pollution and human aggression, *Psych. Rpts.*, *43*, 721–722, 1978.

Latimer, J. A., R. W. Bergstrom, S. R. Hayes, M. K. Liu, J. H. Seinfeld, G. Z. Whitten, M. A. Wojcik, and M. J. Hillyer, *The Development of Mathematical Models for the Prediction of Anthropogenic Visibility Impairment*, EPA Report No. 4503-78-110a,b,c, Environmental Protection Agency, Research Triangle Park, NC, 1978.

Latimer, D. A., T. C. Daniel, and H. Hogo, *Relationship between Air Quality and Human Perception of Scenic Areas*, Publication No. 4323, American Petroleum Institute, Washington, DC, 1980.

Latimer, D. A., H. Hogo, D. H. Hern, and T. C. Daniel, Effects of visual range on the beauty of national parks and wilderness area vistas, in R. D. Rowe and L. G. Chestnut (Eds.), *Managing Air Quality and Scenic Resources at National Parks and Wilderness Areas*, Westview Press, Boulder, Co., 1983.

Loomis, R. J., M. J. Kiphart, D. B. Garnaard, W. C. Malm, and J. V. Molenar, Human perception of visibility impairment, paper presented at the Seventy-Eighth Annual Meeting of the Air Pollution Control Association, Detroit, MI, 1985.

Lowry, E. M., The photometric sensibility of the eye and the precision of photometric observations, *J. Opt. Soc. Am.*, *21*, 32, 1931.

Lowry, E. M., The luminance discrimination of the human eye, *J. Soc. Motion Pictures Television Eng.*, 1951.

Malm, W. C., An examination of the ability of various physical indicators to predict judgment of visual air quality, paper presented at the Seventy-Eighth Annual Meeting of the Air Pollution Control Association, Pittsburgh, PA, 1985.

Malm, W. C., and R. C. Henry, Regulatory perspective of visibility research needs, paper presented at the Eightieth Annual Meeting of the Air Pollution Association, New York, NY, 1987.

Malm, W. C., and M. Pitchford, The use of an atmospheric quadratic detection model to assess change in aerosol concentrations to visibility, paper presented at the Eighty-Second Annual Meeting of the Air Pollution Control Association, Anaheim, CA, 1989.

Malm, W. C., M. Kleine, and K. Kelley, Human perception of visual air quality (layered haze), paper presented at the Conference on Visibility at the Grand Canyon, AZ, 1980.

Malm, W. C., K. Kelley, J. Molenar, and T. Daniel, Human perception of visual air quality (uniform haze), *Atmos. Environ.*, *15*, 1875, 1981.

Malm, W. C., P. Bell, and G. E. McGlothin, Field testing: A methodology for assessing the importance of good visual air quality, in *Proceedings of the Seventy-Seventh Annual Meeting of the Air Pollution Control Association*, Pittsburgh, PA, 1984.

Malm, W. C., D. M. Ross, R. Loomis, J. Molenar, and H. Iyer, An examination of the ability of various physical indicators to predict perception thresholds of plumes as a function of their size and intensity, in P. J. Bhardwaja (Ed.), *Visibility Protection Research and Policy Aspects*, Air Pollution Control Association. Pittsburgh, PA, 1987.

Middleton, P., T. R. Stewart, D. Ely, and C. W. Lewis, Physical and chemical indicators of urban visual air quality, *Atmos. Environ.*, *18*, 861–870, 1984.

Mie, G. *Ann. Phy. Bd. 2*, *25*, *IV*, Fogle, Netherlands, 1908.

Nixon, J. K., Absorption coefficient on NO_2 in the visible spectrum, *J. Chem. Phys.*, *8*, 157, 1940.

Pitchford, M. L., and W. C. Malm, Development and applications of a standard visual index, *Atmos. Environ.*, *28*, 1049–1054.

Ross, D. M., W. C. Malm, and R. J. Loomis, The psychological variation of good visual air quality by national park visitors, paper presented at the Seventy-Eighth Annual Meeting of the Air Pollution Control Association, Detroit, MI, 1985.

Ross, D. M., W. C. Malm, and R. J. Loomis, An examination of the relative importance of park attributes at several national parks, in P. S. Bhardwaja (Ed.), *Transactions of Visibility Protection: Research and Policy Aspects*, Air Pollution Control Association, Pittsburgh, PA, 1987.

Ross, D. M., W. C. Malm, H. K. Iyer, and R. J. Loomis, Human detection of layered haze using natural scene slides with a signal detection paradigm, paper presented at the Eighty-First Annual Meeting of the Air Pollution Control Association, Dallas, TX, Malm, 1988.

Ross, D. M., W. C. Halm, H. K. Iyer, and R. J. Loomis, Human visual sensitivity to layered haze using computer generated images, in C. V. Mathai (Ed.), *Transactions of Visibility and Fine Particles, Air and Waste Management Association*, 582–595, Pittsburgh, PA, 1990.

Ross, D. M., W. C., Malm, and H. K. Iyer, Human visual sensitivity to plumes with a Gaussian luminance distribution: Experiments to develop an empirical probability of detection model, *J. Air Waste Mgmt. Assoc.*, *47*, 370–382, 1997.

Rotton, J., and J. Frey, Atmospheric conditions, seasonal trends, and psychiatric emergencies, in *Replications and Extensions*, American Psychological Association, Washington, DC, 1982.

Rotton, J., T. Barry, M. Milligan, and M. Fitzpatrick, The air pollution experience and interpersonal aggression, *J. Appl. Psych. 9*, 397–412, 1979.

Stewart, T. R., P. Middleton, and D. W. Ely, Urban visual air quality judgements: Reliability and validity, *J. Environ. Psych.*, *3*, 129–145, 1983.

Stewart, T. R., P. Middleton, M. Downton, and D. Ely, Judgements of photographs versus field observations in studies of perception and judgment of the visual environment, *J. Environ. Psych.*, *4*, 283–302, 1984.

vandeHulst, H. C., *Light Scattering by Small Particles*, Dover, New York, 1981.

Ziedner, M., and M. Shechter, Psychological responses to air pollution: Some personality and demographic correlates, *J. Environ. Psych. 8*, 191–208, 1988.

CHAPTER 17

CLOUD CHEMISTRY

STEPHEN E. SCHWARTZ

1 INTRODUCTION

The term *cloud chemistry* is considered here to comprise both cloud composition and reactions that take place in clouds. Clouds are a very special subset of the atmosphere because they present substantial amounts of condensed-phase water (liquid or solid) that can dissolve gases that would otherwise be present in the gas phase, and, as a consequence of condensed-phase reactions, permit reactions to occur that would not otherwise occur or would be much slower. In this sense clouds may be considered to serve as catalysts of atmospheric reactions.

The uptake and reaction of material in clouds, especially sulfur and nitrogen oxides and acids, has received particular attention in the context of gaining improved understanding of the processes responsible for acid deposition. Consequently, the examples developed here focus on these chemical systems. However, much of the resulting understanding of these phenomena is applicable more generally to other systems.

2 CLOUD PHYSICAL PROPERTIES PERTINENT TO CLOUD CHEMISTRY

Clouds consist of a suspension of liquid or solid (ice) particles in air. Thus, formally, a cloud is an aerosol, a suspension of particles in air. However, it is useful to distinguish clouds from clear-air (noncloud) aerosols. The cloud environment is slightly supersaturated with respect to liquid water or ice, respectively. The typical amount of condensed-phase water is 0.1 to 1 g/m^3 (roughly equivalent to 0.1 to 1 l/kg

This chapter has been co-authored under Contract No. DE-AC02-98CH10886 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

of air). The amount of condensed-phase water is substantially lower in cirrus clouds and in polar stratospheric clouds. For condensed-phase amounts substantially exceeding 1 g/m^3, coagulation occurs and precipitation rapidly develops, removing condensed-phase water from the cloud.

A liquid water content of 1 g/m^3 corresponds (within the approximation that the density of water is 1 kg/m^3) to a liquid water volume fraction $L = 1 \times 10^{-6}$, or one part per million by volume. On dimensional grounds the separation between cloud droplets is $\sim L^{-3}$ times the diameter of the droplets; for $L = 1 \times 10^{-6}$, the average interdrop separation is $\sim$100 times the drop diameter. Thus clouds must be considered a sparse suspension of condensed-phase water. Clouds are mostly air. Thus any consideration of cloud chemistry must deal with both the gas phase and the condensed phase.

Despite this sparseness, clouds still contain much more condensed-phase material than cloud-free air. Consider a clear-air aerosol of mass loading of 100 μg/m^{-3}; within the approximation of density equal to 1 kg/m^3, the corresponding condensed-phase volume fraction is 1×10^{-10}. The much greater mass loading of a cloud leads among other things to its greater light scattering, the most distinguishing feature of clouds.

Clouds form when air, containing water vapor, is cooled to a temperature below its dew point. Typically this occurs when air is lifted, for example, buoyant rise of a convective parcel, or larger scale gentle upward motion of warm air over denser cooler air. Cooling by conduction can also be important, for example, in ground fogs, as can radiative cooling. The condensation process defines the number concentration of cloud droplets by activating a certain fraction of preexisting aerosol particles into cloud droplets (see Chapter 19). The number concentration is typically 100 to 1000/cm^3 or 10^8 to 10^9/m^3. Thus within the cloud the condensed-phase water is finely suspended. For droplet concentration of 1×10^9/m^3 and liquid water volume fraction of 1×10^{-6} m^3/m^3, the corresponding volume of an individual droplet is 1×10^{-15} m^3 and the corresponding diameter $\sim 1 \times 10^9$ m or 10 μm.

Invariably there is a dispersion in the diameter of drops; that is, there is a spectrum of cloud droplet sizes. This influences mass transport processes, which are faster for smaller droplets, affecting uptake and reaction of gases in clouds. Typically cloud droplet distributions are rather sharply peaked. This is a consequence of the fact that mass transport of condensing water is faster for smaller droplets thereby allowing the smaller droplets, to "catch up" with the larger ones early in the cloud formation process.

Clouds persist in the atmosphere for a few tens of minutes (short-lived cumulus) to a few tens of hours (persistent stratus). Most clouds evaporate, rather than precipitate, thereby returning dissolved nonvolatile material to the clear air as aerosol particles.

3 SOURCES OF CLOUDWATER COMPOSITION

Cloudwater composition is very much a function of location, being dominated by availability of soluble ionic species. Principal ionic species present in cloudwater

include sodium and chloride, from seawater, sulfate and nitrate anions, and ammonium and hydrogen ion as cations. In regions influenced by industrial emissions of sulfur and nitrogen oxides, cloudwater concentrations of H^+ are commonly 10^{-4} mol/L (molar, M) and not uncommonly 10^{-3} M or higher (Daum et al., 1984).

The fact that cloud droplets form on existing aerosol particles has immediate implications for cloudwater composition. Consider an ammonium sulfate aerosol particle of dry diameter 0.1 μm that serves as a nucleus of a cloud droplet of 10 μm diameter. The volume of the particle is $\sim 10^{-21}$ m^3. For density ~ 1000 kg/m^{-3} and molecular weight 100 g/mol (~ 0.1 kg/mol), the amount of ammonium sulfate contained in the particle is 10^{-17} mol. For this material dissolved in a 10-μm droplet ($\sim 10^{-15}$ m^3) the solution concentration is $\sim 10^{-2}$ mol/m^{-3} or $\sim 10^{-5}$ M. This concentration is at the low end of the range of concentrations of sulfate in cloudwater (and also in precipitation) in regions influenced by industrial emissions (see Chapter 15). It should be stressed that this figure varies as the third power of the particle diameter, that is, an order of magnitude for a factor of 2 in particle diameter. Thus for the particle diameter 0.2 μm, the concentration is 10^{-4} M.

Consider the correspondence between aqueous-phase concentration and the equivalent mixing ratio of the material in air. For one thousand 0.1-μm-diameter, unit-density particles per cm^3, the corresponding mass loading is 1 $\mu g/m^{-3}$, a loading that is rather low in the context of industrialized regions (See Chapter 16), albeit still substantially greater than that characteristic of regions remote from industrial sources. For molecular weight 100, this corresponds to a molar mixing ratio relative to air, $x \approx 0.3$ nmol/mol(air) (ppb). For a substance S that dissolves entirely in cloudwater, the relation between the mixing ratio of the substance x_S in air and concentration in cloudwater is

$$[S] = x_S p_{atm}/LR_g T$$

where [S] is aqueous concentration, p_{atm} is the atmospheric pressure, R_g is the universal gas constant, and T is the absolute temperature. In SI units p_{atm} is in units of pascal and $R_g = 8.3$ J/mol/K. The resulting concentration [S] is in units moles per cubic meters. In practical units (concentration in mol/L and pressure in bar; 1 bar $= 10^5$ Pa)

$$[S](mol/L) = 10^2 x_S p_{atm}(bar)/LR_g T$$

In general the fractional uptake of soluble (ionic) aerosol species into cloudwater is fairly high, approaching unity at low aerosol loading and/or high updraft velocities leading to fairly high maximum supersaturation governing activation of aerosol particles (Leaitch et al., 1996). However, in the case of gases the uptake varies substantially depending on the solubility and/or reactivity of the gas in question.

4 UPTAKE OF GASES INTO CLOUDWATER

In general, a gaseous substance does not dissolve entirely in cloudwater in view of the rather limited solubility of most atmospheric gases in water—if the gas were highly soluble in cloudwater, it would be rapidly rained out and no longer in the atmosphere. The equilibrium concentration of a gaseous substance, S, physically dissolved in a liquid, is given by Henry's law (See Chapter 19).

$$[\mathrm{S(aq)}] = H_S p_S = H_S x_S p_{atm}$$

where H_S is the Henry's law solubility coefficient of the gas. (In practical units, p_S in bar and [S(aq)] in mol/L, i.e., M, H_S has units M/bar.) Abundance of a gas-phase species is expressed in terms of the molar mixing ratio in air x, which is applicable equivalently to substances in gas, aerosol, or solution phases (Schwartz and Warneck, 1995). Characterization of the Henry's law solubility is the first step to understanding the uptake and reaction of a gas in cloudwater. Henry's law solubility coefficients of many gases of atmospheric importance are given in Figure 1.

The ratio of the amount of material in solution to gas phase (distribution ratio), under assumption of Henry's law equilibrium, is given by

$$D_{aq/g} \equiv \frac{\text{moles in aqueous phase}}{\text{moles in gas phase}} = 10^{-2} LH(\mathrm{M/bar}) R_g T$$

If this is written as $D_{aq/g} = H/H_{1/2}$ where $H_{1/2} = 10^{-2}\, LR_g T$, then for any specified value of L, the value of Henry's law solubility coefficient for which the gas is equally distributed between the gas phase and cloudwater is given by $H_{1/2}$. Consider a cloud of rather high liquid volume fraction $L = 10^{-6}$ (i.e., $\sim$1 g/m^{-3} liquid water content); the corresponding value of $H_{1/2}$ is $\sim 4 \times 10^4$ M/bar; $H_{1/2}$ would be correspondingly higher for lower values of L. Comparison with the values of Henry's law solubility coefficients given in Figure 1 shows that virtually all such coefficients are orders of magnitude less than this value, supporting the assertion that reaction of the dissolved gas is required for substantial uptake into cloudwater.

In the case of gases that undergo rapid reversible reaction with water, for example, hydration or acid dissociation, it is necessary to consider the overall solubility equilibrium, not just the Henry's law equilibrium. Consider the solubility equilibrium for the dissolution of an acidic gas, for example, formic acid, HCOOH. The overall equilibrium for this dissolution may be thought to consist of the following steps:

$$\mathrm{HCOOH(g)} = \mathrm{HCOOH(aq)}$$
$$\mathrm{HCOOH(aq)} = \mathrm{H^+(aq)} + \mathrm{COOH^-(aq)}$$

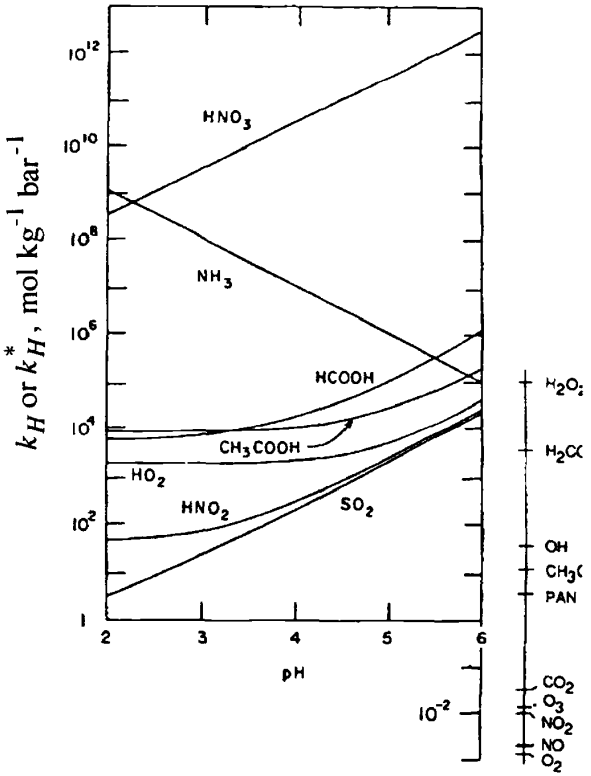


Figure 1 pH dependence of the effective Henry's law coefficient for gases that undergo rapid acid–base dissociation reactions in dilute aqueous solution, as a function of solution pH. Buffer capacity of solution is assumed to greatly exceed incremental concentration from uptake of indicated gas. Also indicated at the right of the figure are Henry's law coefficients for nondissociative gases. $T \sim 300$ K. Modified from Schwartz (1986a).

These reactions sum to give the overall reaction

$$\mathrm{HCOOH(g)} = \mathrm{H^+(aq)} + \mathrm{COOH^-(aq)}$$

The corresponding equilibrium expressions are

$$H_{\mathrm{HCOOH}} = \frac{[\mathrm{HCOOH(aq)}]}{x_{\mathrm{HCOOH}} p_{\mathrm{atm}}} \qquad K_a = \frac{[\mathrm{H^+}][\mathrm{COOH^-}]}{[\mathrm{HCOOH(aq)}]} \qquad K_{\mathrm{eq}} = \frac{[\mathrm{H+}][\mathrm{COOH^-}]}{x_{\mathrm{HCOOH}} p_{\mathrm{atm}}}$$

where K_a is the acid dissociation constant of aqueous formic acid. Depending on the situation, it may be more useful to deal with the overall solubility or with the individual equilibria.

The total concentration of the dissolved gas can be written (here staying with the example of formic acid) as

$$[\text{Formic acid}] \equiv [\text{HCOOH}] + [\text{COOH}^-] = H_{\text{HCOOH}} x_{\text{HCOOH}} p_{\text{atm}} \left(1 + \frac{K_{\text{eq}}}{[\text{H}^+]}\right)$$

It is often a good assumption that the cloudwater is well buffered against change in acid concentration $[\text{H}^+]$ resulting from the incremental uptake of gases present at low partial pressures characteristic of the ambient atmosphere. Under this assumption, $[\text{H}^+]$ is a constant and hence the aqueous concentration is linear in gas-phase partial pressure with an effective Henry's law solubility coefficient defined as:

$$H^*_{\text{HCOOH}} \equiv H_{\text{HCOOH}} \left(1 + \frac{K_{\text{eq}}}{[H^+]}\right)$$

so that one obtains a Henry's law-like expression for the overall solubility,

$$[\text{Formic acid}] = H^*_{\text{HCOOH}} x_{\text{HCOOH}} p_{\text{atm}}$$

In the case of SO_2 there are two acid dissociation equilibria. The effective Henry's law solubility coefficient for S(IV) (the Roman numeral IV denotes the oxidation state) is

$$H^*_{\text{S(IV)}} \equiv H_{\text{SO}_2} \left(1 + \frac{K_{\text{a1}}}{[H^+]} + \frac{K_{\text{a1}} K_{\text{a1}}}{[\text{H}^+]^2}\right)$$

where K_{a1} and K_{a2} denote the first and second dissociation constants, respectively.

Values of effective Henry's law solubility coefficients are shown in Figure 1 as a function of solution pH for the range of pH values typical of cloudwater. The effective solubility coefficient can greatly exceed the Henry's law coefficient for physical dissolution, especially for strong acids, such as nitric acid, and also for ammonia, which is highly soluble in the form of ammonium ion NH_4^+. These effective Henry's law solubility coefficients can also substantially exceed $H_{1/2}$, indicating that at equilibrium, such highly soluble gases as HNO_3 are essentially entirely taken up by cloudwater.

Because the chemical kinetics of acid dissociation reactions are generally quite rapid, the uptake of acidic gases such as HNO_3 is itself quite rapid, under control of mass transport processes rather than chemical kinetics. The mass transport processes governing this uptake are essentially identical to those governing the transfer of water vapor itself to and from cloud droplets, and the solubility of a gas such as HNO_3 is such that the uptake of soluble gases occurs on the time scale of cloud droplet activation and growth, that is taking place on a time scale of a few seconds to

a few tens of seconds. This can result in such soluble gases being preferentially concentrated in the initially formed drops rather than being distributed uniformly throughout the cloud droplet spectrum (Wurzler et al., 1995); this can influence subsequent uptake and reaction of less soluble gases such as SO_2. A gas such as HNO_3 that dissolves in a growing cloud droplet contributes soluble material to the droplet, thereby adding to the Raoult effect of the solute already serving as the cloud condensation nucleus and increasing its cloud nucleating potential. This can have a further influence on cloud droplet composition and can also lead to situations of free cloud droplet growth at relative humidity slightly below 100% (Kulmala et al., 1997).

5 REACTIVE UPTAKE OF GASES BY CLOUDWATER

Without further reaction the fractional uptake of SO_2 into cloudwater is low, even at fairly high pH. The same is true *a fortiori* for NO_2, which does not undergo acid dissociation reaction in aqueous solution. However, there is a strong thermodynamic driving force in clouds for the reactive uptake of these gases to form sulfuric and nitric acids, respectively, the principal species contributing to acid deposition. This situation has stimulated substantial research interest in the processes whereby these gases are transformed into the acids and incorporated into cloudwater. The present understanding of these reactive uptake processes is that in the case of SO_2, the process consisting of uptake of SO_2 followed by aqueous-phase oxidation contributes substantially to the uptake of sulfuric acid by cloudwater and to the deposition of this material in precipitation. In contrast, the uptake process for NO_2 to form nitric acid appears to be dominated by gas-phase oxidation followed by uptake of the oxidized species. This section presents the formalism by which the rate of aqueous-phase reaction in cloudwater may be evaluated treating these two gases as examples.

Consider the rate of aqueous-phase reaction of dissolved sulfur-IV to be given by

$$\frac{-d[\mathrm{S(IV)}]}{dt} = \frac{d[\mathrm{S(VI)}]}{dt} = k^{(1)}[\mathrm{S(IV)}]$$

where $k^{(1)}$ denotes an effective first-order rate coefficient, which in general may be equal to a second-order rate coefficient times the concentration of a second reagent. The reaction need not be first-order in the reagent sulfur-IV; additional power(s) of [S(IV)] could be incorporated within the effective first-order rate coefficient $k^{(1)}$. It is useful to refer the reaction rate to the total S(IV) concentration because of equilibration of individual sulfur-IV species within solution, $SO_2(aq)$, HSO_3^- or bisulfite, and SO_3^{2-} or sulfite, that is rapid relative to depletion by reaction. The aqueous-phase reaction rate can be related to the gas-phase mixing ratio of SO_2 by solubility equilibria between aqueous-phase concentration of S(IV) and gas-phase partial pressure of SO_2, under assumption that these equilibria apply. This phase equilibrium is expected to hold if mass transport rates coupling the two phases are sufficiently fast

to replenish the aqueous-phase material that is depleted by reaction, a situation that is normally expected to obtain, as discussed below. Hence

$$-\frac{d[\mathrm{S(IV)}]}{dt} = \frac{d[\mathrm{S(VI)}]}{dt} = k^{(1)} H^*_{\mathrm{S(IV)}} x_{\mathrm{SO_2}} p_{\mathrm{atm}}$$

Under assumption that the aqueous-phase rate is uniform within a given region of a cloud, then the rate of reaction, expressed as a rate of decrease in the mixing ratio of SO_2, is

$$\frac{dx_{\mathrm{SO_2}}}{dt} = -\{10^{-2} L R_g T k^{(1)} H^*_{\mathrm{S(IV)}}\} x_{\mathrm{SO_2}} = -k^{(1)}_{\mathrm{eff}} x_{\mathrm{SO_2}}$$

The quantity in braces is an effective first-order rate coefficient of aqueous-phase reaction, referred to the gas-phase mixing ratio; this quantity may be directly employed in evaluating rates of reactions or in comparison to rate coefficients for loss by gas-phase reactions. Note that $k^{(1)}_{\mathrm{eff}}$ scales linearly with liquid water content and with Henry's law solubility coefficient. The more water present to serve as volume of reactor, the faster the reaction. Likewise, the more soluble the reagent gas, the faster the reaction. Evaluation of the rate of a specific reaction requires knowledge of the effective first-order rate coefficient of aqueous-phase reaction, $k^{(1)}$. For this, one must identify the mechanism and rate of aqueous-phase reaction.

There is a strong thermodynamic driving force for oxidation of dissolved SO_2 by molecular oxygen, which, because of its abundance, might be thought to be the key oxidant of SO_2 in cloudwater. However, this reaction is quite slow unless catalyzed, for example, by transition metal ions. Although catalyzed oxidation of dissolved sulfur-IV by dissolved molecular oxygen may be of some importance in some circumstances, the species that have been identified as of principal importance in oxidation of sulfur-IV in cloudwater are the strong oxidants ozone (O_3) and hydrogen peroxide (H_2O_2). Ozone is commonly present in the atmosphere at a mixing ratio of 30 to 50 nmol/mol. Hydrogen peroxide is present at much lower abundance, ~1 nmol/mol. These mixing ratios compare with those for SO_2 of order 10 nmol/mol in regions influenced by industrial emissions, to much lower at locations well removed from sources.

Consider first the ozone reaction. The rate of aqueous-phase reaction is given as

$$-\frac{d[\mathrm{S(IV)}]}{dt} = \frac{d[\mathrm{S(VI)}]}{dt} = k^{(2)}[\mathrm{S(IV)}][\mathrm{O_3}], \qquad \text{i.e., } k^{(1)} = k^{(2)}[\mathrm{O_3}]$$

where $k^{(2)}$ is a second-order rate constant that must be determined by laboratory measurement and has been found to exhibit a strong pH dependence, increasing with increasing pH (Fig. 2a). The concentration of dissolved ozone is related to the gas-

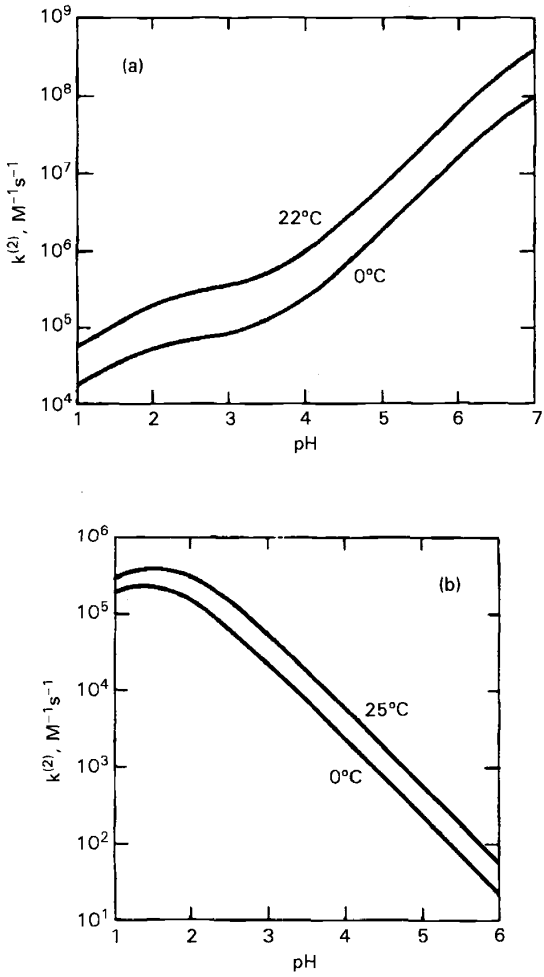


Figure 2 Effective second-order rate coefficients for aqueous-phase reaction of S(IV) with O_3 (a) and with H_2O_2 (b) as a function of pH. Modified from Schwartz (1988).

phase mixing ratio of this species again under assumption of solubility equilibrium, as

$$[O_3] = H_{O_3} x_{O_3} p_{atm}$$

so that

$$k_{eff}^{(1)} = 10^{-2} L R_g T k^{(2)} H_{O_3} x_{O_3} H^*_{S(IV))} p_{atm}$$

Combining the kinetic and solubility data permits the rate of reaction to be evaluated for known or assumed conditions of cloud liquid water content and partial pressures of reagent gases (Fig. 3). Here the left-hand ordinate gives the rate of aqueous-phase reaction. The right-hand ordinate gives the effective first-order rate coefficient of aqueous-phase reaction, referred to the gas-phase mixing ratio; $k_{eff}^{(1)}$ for indicated conditions, expressed in units of percent per hour. Note the strong pH dependence, rate increasing with pH, resulting from the pH dependences of sulfur-IV solubility (Fig. 1) and kinetic rate constant (Fig. 2). The ozone reaction is quite rapid at high pH. However, because of production of sulfuric acid as the reaction

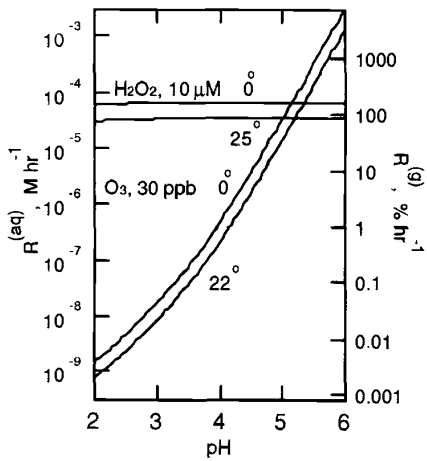


Figure 3 Instantaneous rate of aqueous-phase oxidation of S(IV) by H_2O_2 and O_3, evaluated as a function of pH for representative nonurban reagent concentrations. The rates scale approximately linearly with reagent concentrations. The right-hand ordinate gives the oxidation rate of SO_2 referred to the gas-phase partial pressure and expressed as percent per hour for a liquid water content $L = 1 \times 10^{-6}$ (1 cm^3/m^{-3}), the rate scales approximately linearly with L. For the H_2O_2 reaction the indicated aqueous-phase concentration of H_2O_2 corresponds to total mixing ratio of this species (gas plus aqueous phase; the two are comparable) of $\sim 0.6 \times 10^{-9}$. See ftp site for color image.

proceeds, the pH rapidly becomes lower, decreasing the rate. Although a strong acid concentration of 10 μM (pH 5) is quickly reached, in perhaps 10 min, a much greater time, ~10 h, is required to reach an acid concentration of 50 μM. For this reason the ozone reaction is unlikely to account for cloudwater acidities of 10^{-4} to 10^{-3} M commonly observed in regions influenced by industrial emissions of SO_2.

Now consider the hydrogen peroxide reaction, whose aqueous-phase rate is given by:

$$-\frac{d[\mathrm{S(IV)}]}{dt} = \frac{d[\mathrm{S(VI)}]}{dt} = k^{(2)}[\mathrm{S(IV)}][\mathrm{H_2O_2}]$$

This reaction is acid catalyzed; that is, the second-order aqueous-phase rate constant increases with decreasing pH (Fig. 2*b*). The pH dependences of solubility and reaction kinetics now cancel to yield a reaction rate that is roughly independent of pH throughout the pH range pertinent to cloudwater acidification (Fig. 3). For the conditions given in Figure 3 the rates of the O_3 and H_2O_2 reactions are equal roughly at pH 5. An effective first-order reaction rate of 100%/h corresponds to a $1/e$ lifetime of SO_2 of 1 h; the actual lifetime would depend on actual conditions. The H_2O_2 reaction is the only identified atmospheric reaction capable of maintaining the SO_2 oxidation rate sufficiently rapid to produce observed cloudwater H^+ and SO_4^{2-} concentrations on time scales pertinent to cloud acidification.

In the case of the ozone reaction, ambient mixing ratios of O_3 are generally sufficiently in excess of those of SO_2 that depletion of O_3 need not be considered. However, ambient concentrations of H_2O_2, which are typically below 3 nmol/mol are often much less than ambient SO_2 mixing ratios. This leads to a situation where the reaction proceeds rapidly to completion by exhausting the H_2O_2 reagent. If on the other hand SO_2 mixing ratios are the lesser, then the reaction can rapidly and completely exhaust ambient SO_2. The time scale of this process, a few tens of minutes for representative mixing ratios in the nmol/mol region, leads to a situation where the extent of reaction is controlled by the limiting reagent. Such appears to be the case as indicated by field measurements simultaneously examining H_2O_2 and SO_2 mixing ratios in clouds. A survey of nonprecipitating stratiform clouds indicated that although either species is frequently present at nmol/mol mixing ratios, appreciable mixing ratios of the two species are virtually never simultaneously present, (Daum, 1990). The contribution of this reaction to cloudwater acidification has been directly confirmed by field measurements under well-defined flow conditions, including experiments with artificially introduced SO_2 and inert tracers, showing concomitant decreases in SO_2, and H_2O_2 and increases in H^+ and SO_4^{2-} consistent with this reaction. This reaction is now thought to be the major contributor to atmospheric oxidation of SO_2, contributing both to acid precipitation and, in the likely event of cloud evaporation to sulfate aerosol, a principal component of atmospheric aerosols.

Based on laboratory and industrial experience nitrogen dioxide (NO_2) is known to be highly reactive with liquid water forming nitric and/or nitrous acids, the initial reaction being

$$2NO_2(N_2O_4) + H_2O \rightarrow 2H^+ + NO_3^- + NO_2^-$$

for which there is a strong thermochemical driving force; the N_2O_4 in parentheses indicates the possible participation of the NO_2 dimer, dinitrogen tetroxide. (This reaction is the basis for industrial manufacture of nitric acid.) It was therefore assumed by many atmospheric chemists that NO_2 would be rapidly taken up in cloudwater in the ambient atmosphere. Consideration of the mechanism of this reaction gives the aqueous-phase rate expression,

$$-\frac{d[NO_2]}{dt} = 2k^{(2)}H^2_{NO_2}x^2_{NO_2}p^2_{atm}$$

Determination of the Henry's law coefficient for NO_2 and the second-order reaction rate constant permitted evaluation of this rate for atmospheric conditions. Such evaluations have indicated that this rate is much too slow to contribute appreciably to NO_2 uptake by cloudwater at ambient concentrations. The reason for this, and for the great difference with experience at high NO_2 concentrations, is that the reaction is second-order in the concentration of a very weakly soluble gas. Comparisons of NO_2 mixing ratios in clouds with those in clear air in the vicinity of clouds indicates that the fractional uptake of NO_2 into cloudwater is quite small, lending confirmation to the above picture. Alternative possible mechanisms for NO_2 uptake include reaction with reducing species dissolved in cloudwater, as NO_2 is a fairly strong oxidant.

An alternative mechanism that may be important, especially at night, is initiated by the gas-phase reactions

$$NO_2 + O_3 \rightarrow NO_3 + O_2$$

followed by

$$NO_3 + NO_2 \rightarrow N_2O_5$$

with N_2O_5 being taken up by cloudwater by reaction to form nitric acid:

$$N_2O_5 + H_2O(l) \rightarrow 2H^+ + 2NO_3^-$$

The rate of reaction is controlled by the rate of the initiation reaction of NO_2 with O_3, which is several percent per hour at typical ozone mixing ratios of a few tens

of nmol/mol. The reason that this appears to be important at night but not during the day is that photolysis of NO_3 by visible radiation

$$NO_3 + h\nu \rightarrow NO + O_2$$

is the major sink of NO_3 during the day, thereby cutting off the overall reaction.

In addition to these acidification reactions several other in-cloud reactions have been identified as of importance or potential importance in atmospheric chemistry. The hydroperoxy radical, HO_2, which plays an important role in gas-phase photochemistry as part of the chain of reactions leading to ozone formation by oxidation of NO and hydrocarbons, is thought to be rather soluble in water because of its weak-acid dissociation:

$$HO_2 \rightarrow H^+ + O_2^-$$

The dissolved material undergoes rapid self-reaction to form hydrogen peroxide. It has been suggested that the occurrence of this process can substantially influence the ozone budget in the remote troposphere. However, the process remains somewhat speculative in view of the lack of firm information on the solubility of the HO_2 radical.

Several studies have demonstrated substantial aqueous-phase formation of H_2O_2 by photochemical reactions in collected cloudwater. The exact processes are not yet elucidated but evidently involve trace organic species, which are difficult to characterize. Such reactions may contribute substantially to SO_2 oxidation in situations where this oxidation is limited by the amount of H_2O_2 initially present. More generally, it may be noted that photochemical reactions, in both gas and solution phases, may be enhanced in the tops of clouds because of enhanced photolysis fluxes, by a factor of 5 or more, that result from multiple scattering of solar radiation within clouds.

6 COUPLED MASS TRANSPORT AND CHEMICAL REACTION

As indicated above, quantitative evaluation of the rates of aqueous-phase reactions in clouds are predicated on the assumption that the rate of mass transport processes coupling the gas-phase reservoir of reagent gas to the solution phase within individual cloud droplets is sufficiently fast to maintain the Henry's law equilibrium in competition with the sink of dissolved material by aqueous-phase reaction. The pertinent mass transfer processes are gas-phase diffusion, from the bulk of the gas phase to the gas–liquid interface; transfer across the interface, as governed by the gas-kinetic collision rate and the mass accommodation coefficient, the fraction of collisions resulting in transfer of material across the interface, a property characteristic of individual gases and solutions; and aqueous-phase diffusion of the dissolved gas occurring concomitantly with aqueous-phase reaction. In general, if the reaction is sufficiently slow, mass transport is sufficiently rapid to maintain the solubility

equilibria, but departure from equilibrium occurs for sufficiently rapid reaction rates. Criteria for the onset of this "mass transport limitation" of the rate of aqueous-phase reactions in clouds have been developed in terms of drop radius, Henry's law coefficient, effective first-order reaction rate coefficient, diffusion coefficients, and mass accommodation coefficients. For the most part, the rate of reaction of SO_2 in cloudwater appears only minimally limited by mass transport rates, the exception being the ozone reaction at high pH, under which condition both the solubility and effective first-order rate coefficient are quite large.

7 SUMMARY

Clouds present substantial concentrations of liquid-phase water, which can potentially serve as a medium for dissolution and reaction of atmospheric gases. The important precursors of acid deposition, SO_2, and nitrogen oxides NO and NO_2 are only sparingly soluble in clouds without further oxidation to sulfuric and nitric acids. In the case of SO_2, aqueous-phase reaction with hydrogen peroxide and to lesser extent ozone are identified as important processes leading to this oxidation, and methods have been described by which to evaluate the rates of these reactions. The limited solubility of the nitrogen oxides precludes significant aqueous-phase reaction of these species, but gas-phase reactions in clouds can be important especially at night.

REFERENCES

Daum, P. H., Observations of H_2O_2 and S(IV) in air, cloudwater and precipitation and their implications for the reactive scavenging of SO_2, *Atmos. Res.*, *25*, 89–102, 1990.

Daum, P. H., T. J. Kelly, S. E. Schwartz, and L. Newman, Measurements of the chemical composition of stratiform clouds, *Atmos. Environ.*, *18*, 2671–2684, 1984.

Kulmala, M., A. Laaksonen, R. J. Charlson, and P. Korhonen, Clouds without supersaturation, *Nature*, *388*, 336–337, 1997.

Leaitch. W. R., C. M. Banic, G. A. Isaac, M. D. Couture, P. S. K. Liu, I. Gultepe, S.-M. Li, K. I. Kleinman, P. H. Daum, and J. I. MacPherson, Physical and chemical observations in marine stratus during the 1993 North Atlantic Regional Experiment: Factors controlling cloud droplet number concentrations, *J. Geophys. Res.*, *101*, 29123–29135, 1996.

Schwartz, S. E, Chemical conversions in clouds, in S. D. Lee, T. Schneider, L. D. Grant, and P. J. Verkerk (Eds.), Lewis Publishers, Chelsea, MI, 1986a, pp. 349–375.

Schwartz, S. E., Mass-transport considerations pertinent to aqueous-phase reactions of gases in liquid-water clouds, in, W. Jaeschke (Ed.), *Chemistry of Multiphase Atmospheric Systems*, Springer, Heidelberg, 1986b, pp. 415–471.

Schwartz, S. E., Mass-transport limitation to the rate of in-cloud oxidation of SO_2: Re-examination in the light of new data, *Atmos. Environ.*, *22*, 2491–2499, 1988.

Schwartz, S. E., and P. Warneck, Units for use in atmospheric chemistry, *Pure Appl. Chem. 67*, 1377–1406, 1995.

Wurzler S., A. I. Flossmann, H. R. Pruppacher, and S. E. Schwartz, The scavenging of nitrate by clouds and precipitation. I. A theoretical study of the uptake and redistribution of $NaNO_3$ particles and HNO_3 gas by growing cloud drops using an entraining air parcel model, *J. Atmos. Chem.*, *20*, 259–280, 1995.

CHAPTER 18

DRY DEPOSITION

M. L. WESELY

1 INTRODUCTION

Dry deposition refers to the removal of trace substances from the atmosphere without the aid of precipitation. Dry deposition cleanses the air and delivers substances to the Earth's surface. In air, the vertical transfer is accomplished primarily by turbulent mixing; for particles whose diameters are greater than 1 to 2 μm, vertical transfer is aided by gravitational settling. As air comes into contact with surface elements, gas molecules can react with surface materials or dissolve in them. Particles can be captured by interception or impaction with the surface elements. In comparison to wet deposition, in which substances are carried in precipitation to the surface, dry deposition is a slower, more continuous process and is profoundly affected by the physical, chemical, and biological properties of the surface.

The downward vertical mass flux density (deposition rate per unit area), divided by concentration C at a specified height, is conventionally known as the deposition velocity V_d. The value of this velocity can vary greatly depending on the properties of the substance of interest and local atmospheric and surface conditions. Nevertheless, the concept of a deposition velocity is highly useful in many applications because it produces the deposition rate when multiplied by a measured or modeled concentration. As a general guideline, a deposition velocity of 0.1 cm/s is small, 1.0 cm/s is moderately large, and several centimeters per second is near the limit that is physically possible on the basis of turbulent mixing alone. The corresponding residence times of the constituents in the lower atmosphere can be weeks, days, or hours, respectively, when they are controlled only by dry deposition. Relatively inert gases with very small deposition velocities can have lifetimes of several years in the atmosphere if other sources of removal, such as chemical transformations, are weak.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

Because dry deposition is a surface process, it can be treated as a boundary condition in models that compute atmospheric chemical budgets. In numerical models, the deposition rate is usually estimated on the basis of deposition velocity fairly close to the surface, typically at a height less than 50 m, with the assumption that the vertical flux below such small heights does not change substantially with height. This assumption is valid if the lower atmosphere has concentrations and mixing properties that are horizontally uniform and steady with time over the period of a few hours. Another requirement for nearly constant fluxes with height is that the substance of interest does not undergo chemical and physical changes that are rapid in comparison to the time scales of local vertical mixing. When these requirements are met, the vertical flux can be estimated as

$$F = -V_d C \tag{1}$$

where we have adopted the convention that a flux directed downward is negative. This type of formulation is intended only for the case of a flux being directed downward, when V_d is positive; it has little merit for substances that are emitted from the surface because ambient concentration often has little effect on emission rates.

This chapter provides a brief, somewhat introductory, overview of dry deposition. The reader can find considerable additional information in the scientific literature. For example, a review of the state of the science after considerable research on acidic deposition during the 1980s in the United States was provided by Hicks et al. (1989), and measurement techniques were reviewed by Businger (1986). Some European perspectives on acidic deposition were provided by Erisman and Draaijers (1995), and reviews of the status of dry deposition knowledge were conducted recently by Lovett (1994), Seinfeld and Pandis (1998), and Wesely and Hicks (2000).

2 FORMULATION OF DEPOSITION VELOCITY

A common simple method of evaluating Eq. (1) is by analogy to Ohm's Law, where F corresponds to current, V_d corresponds to the inverse of the total resistance, and C corresponds to the voltage, referenced to electrical ground corresponding to a concentration of zero that is assumed to occur somewhere in the surface. From this analogy, the deposition velocity can be expressed in terms of three resistances in series.

$$V_d = (R_a + R_b + R_c)^{-1} \tag{2}$$

Here, R_a represents the aerodynamic resistance to transfer associated with turbulent mixing above the surface, R_b is the resistance of the quasilaminar sublayer of air in contact with surface elements, and R_c is the bulk resistance of the surface. The values of R_a and R_b can be estimated with readily available micrometeorological formulations. Various schemes exist in the literature for depicting and evaluating

these resistances (e.g., Hicks et al., 1987; Wesely, 1989), and Figure 1 shows a relatively complex version that includes some of the many resistances in series and parallel that can be constructed for R_c.

The existence and the relative importance of each flux pathway shown in Figure 1 tend to be unique for each type of surface and each substance. Each resistance term itself normally must be parameterized in terms of surface properties. In field experiments, R_c is often found as the residual, unmeasured quantity in Eqs. (1) and (2). Some components of R_c are not measured directly but are typically inferred over a

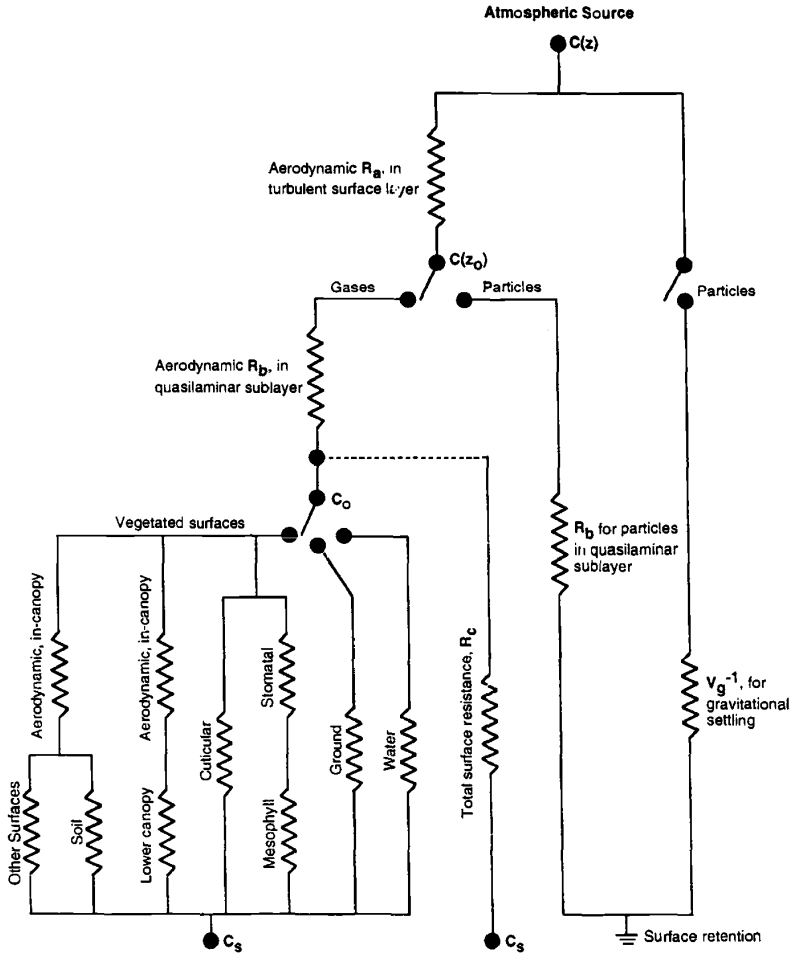


Figure 1 Resistance scheme for dry deposition.

particular surface as environmental conditions change. When leaf stomatal openings close at night, for example, the bulk resistance to deposition on the outer surfaces of leaves and the ground below vegetative canopies can be inferred. The resistance of the waxy cuticle is sometimes measured in the laboratory, and the resistance of the ground surface beneath the canopy is occasionally evaluated with flux measurements there. In the parameterizations that are generated, important variables include environmental factors (such as solar radiation, temperature, air humidity, wetness of the surface caused by dew and rainfall, and soil moisture content) and details of the surface (such as height of vegetative canopy, amount of leaf area, species of vegetation, and soil pH). For bodies of water, the structure and size of waves can be important.

For particle deposition, R_c is not commonly considered explicitly and the deposition velocity is expressed in terms of R_a, R_b, and gravitation settling velocity V_g. However, R_b embodies several somewhat complex processes involving transport through the quasilaminar sublayer, interception of particles by fine elements of the surface, and inertial impaction of particles on the surface. Theoretical formulations for both R_b and V_g usually include a strong dependency on particle size.

Although R_a, R_b, and aerodynamic resistances in canopies are considered separately, they are all strongly affected by turbulence parameters. The turbulent mixing induced by buoyancy forces associated with surface heating by solar radiation can directly alter R_a and in-canopy resistances. The roughness of the surface, a primary factor in evaluating R_a, is linked indirectly to the vegetative properties that affect the resistances of elements of a vegetative canopy. Somewhat more confusing is the fact that in-canopy resistances are implied in Figure 1 to be controlled purely by turbulence, but the distribution of "sinks" in canopies can alter the value of the in-canopy aerodynamic resistance because the latter is actually a composite of vertically distributed air-phase resistances to many surface elements. For this and other reasons noted below, the approach that uses Eq. (2) and Figure 1 is considered by many researchers to be oversimplified.

Several other difficulties exist with the resistance analogy for dry deposition. Some experiments have shown, for example, that weak mixing at night in tall canopies might lead to storage of chemicals in air in the sheltered areas. Deep snowpack might store some relatively insoluble substances. Gusty winds can resuspend particles. Relatively inert gases might temporarily dissolve in surface materials and later be reemitted. Some trace gases are emitted from natural surfaces, sometimes at greater rates when the ambient concentrations are small. To overcome such difficulties, experimental observations of deposition velocities sometimes provide the primary information used to evaluate Eq. (1), or more sophisticated models of air–surface exchange are developed.

3 DEPOSITION VELOCITY ESTIMATES

The great diversity of airborne trace chemical properties, surface conditions, and environmental conditions prevents the generation of universally applicable dry

deposition parameterization schemes. Studies have tended to focus on substances important in atmospheric chemistry or likely to be harmful to human health, biota, and man-made materials. The types of surface most often chosen for investigation are ones that are fairly common in a given region, because the total amount of substances removed from the atmosphere depends on the amount of areal exposure to the surfaces. For example, the deposition of O_3 to agricultural areas and forests in the eastern half of the United States has been studied fairly frequently. Figure 2 shows the results of some studies conducted by Argonne National Laboratory in the 1990s. As can be seen, the deposition velocity to soybean fields tends to be larger than for maize, a pattern related to the structures and physiological properties of the plants. Cloudy conditions and the resulting reduction in solar radiation for the central day in Figure 2 caused leaf stomata to be partially closed and the value of R_c to be fairly large. At night, the stomata were closed for all three canopies, leaving open only the deposition pathway to the outer surfaces of the vegetation and the soil surface beneath. Deposition to the tallgrass prairie tended to be suppressed in general because of the effects of fairly dry soil and the tendency of the grass to increase stomatal resistance in such conditions to reduce transpiration of moisture.

Ozone is taken up through plant stomata because destruction of O_3 occurs fairly rapidly in the substomatal cavities. The oxidization by O_3 of some organic substances in solution in the film of water that envelops the mesophyll cells is the primary reason for the very small mesophyll resistance. Ozone also reacts strongly with many surface materials, although the waxy outer coating of leaves is an effective barrier. Because O_3 is poorly soluble in water, flux pathways are insignificant to water that is free from significant amounts of substances with which O_3 reacts. In

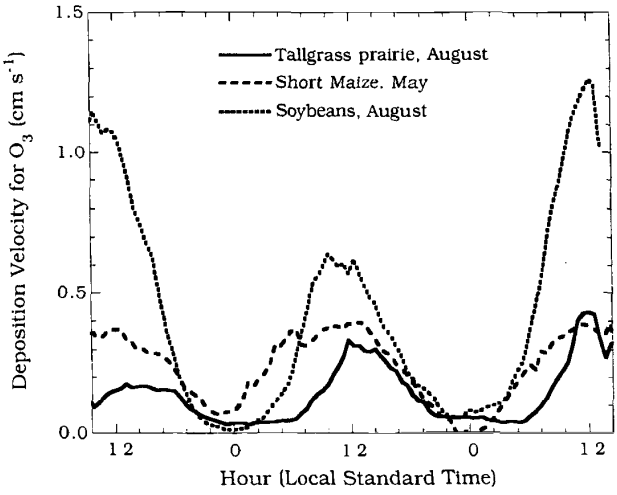


Figure 2 Observations of the dry deposition velocity for ozone over three types of surfaces.

general, measures of the oxidizing capacity of various substances provide a means of evaluating their ability to be destroyed at surfaces of various types.

Many studies have shown that SO_2 is also taken up by vegetation with practically no mesophyll resistance. The primary factor that affects SO_2 removal is its fairly large effective solubility in water. Usually the amount of exposure to water is a major factor in assessing the deposition velocity of substances that dissolve and dissociate rapidly in water. Table 1 shows deposition velocity values for SO_2, O_3, and NO_2 for various surfaces, based on experimental observations and resistance models. As can be seen, the deposition velocities for NO_2 tend to be smaller than for SO_2 and O_3, mainly because the water solubility of NO_2 is small and its ability to oxidize surface materials is weak compared to that of O_3. In general, measures of the effective solubility and oxidizing capacity of gases provide a means of estimating their deposition velocities relative to those seen experimentally for SO_2 and O_3.

TABLE 1 Typical Deposition Velocities (cm/s[1]) for SO_2, O_3, and NO_2 at a Height of 10 m[a]

	Soybeans		Grassland, Maize		Deciduous Forest		Coniferous Forest	
Substance	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*
Midsummer with Lush Vegetation								
SO_2	1.4	0.4	0.8	0.3	0.9	0.1	0.6	0.1
O_3	1.0	0.2	0.7	0.2	0.8	0.1	0.5	0.1
NO_2	0.8	0.1	0.4	0.05	0.7	0.03	0.4	0.03
Autumn with Unharvested Cropland								
SO_2	0.4	0.2	0.4	0.2	0.2	0.1	0.3	0.1
O_3	0.4	0.2	0.4	0.2	0.2	0.1	0.3	0.1
NO_2	0.1	0.1	0.1	0.05	0.05	0.03	0.2	0.03
Late Autumn after Frost, No Snow								
SO_2	0.5	0.2	0.2	0.1	0.1	0.1	0.1	0.1
O_3	0.5	0.2	0.4	0.2	0.2	0.1	0.2	0.1
NO_2	0.1	0.1	0.1	0.05	0.1	0.04	0.1	0.03
Winter, Snow on Ground and Near Freezing								
SO_2	0.5	0.2	0.5	0.2	0.1	0.1	0.3	0.1
O_3	0.1	0.03	0.1	0.03	0.2	0.03	0.1	0.04
NO_2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Transitional Spring with Partially Green Short Annuals								
SO_2	1.0	0.3	0.7	0.2	0.5	0.1	0.3	0.1
O_3	0.7	0.2	0.5	0.2	0.5	0.1	0.3	0.1
NO_2	0.4	0.1	0.2	0.05	0.3	0.04	0.2	0.03

[a]Cases 1 and 2 for each surface type correspond to solar irradiances of 500 and 0 W/m^2, respectively. Dry surfaces and moderate wind speeds are assumed.

The deposition of nonpolar, nonreactive gases such as some organic compounds is usually assumed to be small, although solubility in lipids in vegetation might slightly enhance deposition. Studies have shown that this pathway is measurable but very small. Deposition velocities of less than 0.1 cm/s are likely.

Particle deposition velocities can be strongly dependent on particle size. For particles smaller than 0.1 to 0.2 μm in diameter, deposition by transport through the quasilaminar sublayer can be fairly strong; the extremely fine particles diffuse through air similarly to molecules of gas. Particles larger than 1 to 2 μm are deposited mainly by gravitational settling, for which the associated deposition velocities can be several centimeters per second. For the so-called accumulation size mode, in which particle diameters are larger than 0.1 to 0.2 μm and smaller than 1 to 2 μm, mechanisms of deposition are often thought to be ineffective. Some field studies have shown, however, that processes of interception and impaction in gusty wind conditions can enhance deposition velocities substantially, to values exceeding 0.5 cm/s during daytime conditions over typical terrestrial surfaces. Such deposition velocities have been seen over grass for sulfate, which usually exists primarily in the accumulation size mode; values exceeding 1.0 cm/s for sulfate and nitrate have been seen over partially wetted coniferous forests.

4 MODELS OF DEPOSITION VELOCITY

Models have become significantly more sophisticated during the past two decades and are becoming more effective tools for the environmental worker who must make estimates of deposition rates of trace chemicals. "Big-leaf models" that use Eq. (2) with little breakdown of R_c into component resistances have been supplanted to some extent by multilayer canopy models for vegetated surfaces at specific sites where local conditions are observed directly (e.g., Meyers and Baldocchi, 1988; Meyers et al., 1998). Variations of big-leaf models in which R_c is represented by several possible flux pathways have been used extensively in dry deposition modules intended for regional- and large-scale numerical models of atmospheric chemistry (e.g., Pleim et al., 1984; Wesely, 1989; Padro and Edwards, 1991; Benkovitz et al., 1994; Ganzeveld and Lelieveld, 1995).

The potential is high for advancing the accuracy of dry deposition estimates by using advanced atmospheric models with notably improved descriptions of the surface conditions that affect dry deposition. Third-generation models are expected to have capabilities that will reduce the dependency on empirically derived resistance values and provide a means of coupling deposition and emission more closely (Peters et al., 1995). Third-generation models are also likely to incorporate better simulations of the structure of the planetary boundary layer, to provide estimates of soil moisture content and evapotranspiration that can be valuable inputs to dry deposition modules, and to allow the use of parameterizations of vegetative processes that are based on physiological processes, such as processes that control photosynthesis and uptake of carbon dioxide.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy under contract W-31-109-Eng-38, as part of the Atmospheric Chemistry Program of the Office of Science, Office of Biological and Environmental Research, Environmental Sciences Division.

REFERENCES

Benkovitz, C. M., C. M. Berkowitz, R. C. Easter, S. Nemesure, R. Wagener, and S. E. Schwartz, Sulfate over the North Atlantic and adjacent continental regions: Evaluation for October and November 1986 using a three-dimensional model driven by observation-derived meteorology, *J. Geophys. Res.*, *99*, 20725–20756, 1994.

Businger, J. A., Evaluation of the accuracy with which dry deposition can be measured with current micrometeorological techniques, *J. Climate Appl. Meteorol.*, *25*, 1100–1124, 1986.

Erisman, J. W., and G. P. J. Draaijers, *Atmospheric Deposition in Relation to Acidification and Eutrophication*. Elsevier, New York, 1995.

Ganzeveld, L., and J. Lelieveld, Dry deposition parameterization in a chemistry general circulation model and its influence on the distribution of reactive trace gases, *J. Geophys. Res.*, *100*, 20999–21012, 1995.

Hicks, B. B., D. D. Baldocchi, T. P. Meyers, R. P. Hosker, Jr., and D. R. Matt, A preliminary multiple resistance routine for deriving dry deposition velocities from measured quantities, *Water Air Soil Pollut.*, *36*, 311–330, 1987.

Hicks, B. B., R. R. Draxler, D. L. Albritton, F. C. Fehsenfeld, J. M. Hales, T. P. Meyers, R. L. Vong, M. Dodge, S. E. Schwartz, R. L. Tanner, C. I. Davidson, S. E. Lindberg, and M. L. Wesely, *Atmospheric Processes Research and Process Model Development*, State of the Science/Technology, Report No. 2, National Acid Precipitation Assessment Program, Superintendent of Documents, Government Printing Office, Washington, D. C., 1989.

Lovett, G. M., Atmospheric deposition of nutrients and pollutants in North America: An ecological perspective, *Ecol. Appl.*, *4*, 629–650, 1994.

Meyers, T. P., and D. D. Baldocchi, A comparison of models for deriving dry deposition fluxes of O_3 and SO_2 to a forest canopy, *Tellus*, *40B*, 270–284, 1988.

Meyers, T. P., P. Finkelstein, J. Clarke, T. G. Ellestad, and P. F. Sims, A multilayer model for inferring dry deposition using standard meteorological measurements, *J. Geophys. Res.*, *103*, 22645–22661, 1998.

Padro, J., and G. C. Edwards, Sensitivity of ADOM dry deposition velocities to input parameters: A comparison with measurements for SO_2 and NO_2 over three land-use types, *Atmos.-Ocean.*, *29*, 667–685, 1991.

Peters, L. K., C. M. Berkowitz, G. R. Carmichael, R. C. Easter, G. Fairweather, S. J. Ghan, J. M. Hales, L. R. Leung, W. R. Pennell, F. A. Potra, R. D. Saylor, and T. T. Tsang, The current status and future direction of Eulerian models in simulating the tropospheric chemistry and transport of trace species: A review, *Atmos. Environ.*, *29*, 189–222, 1995.

Pleim, J. E., A. Venkatram, and R. Yamartino, *ADOM/TADAP Model Development Program*, Vol. 4: *The Dry Deposition Module*, Ontario Ministry of the Environment, Rexdale, Canada, 1984.

Seinfeld, J. H., and S. N. Pandis, *Atmospheric Chemistry and Physics* Wiley, New York, 1998.

Wesely, M. L., Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models, *Atmos. Environ.*, *23*, 1293–1304, 1989.

Wesely, M. L., and B. B. Hicks, A review of the current status of knowledge on dry deposition, *Atmos. Environ.*, *34*, 2261–2282, 2000.

CHAPTER 19

FATE OF ATMOSPHERIC TRACE GASES: WET DEPOSITION

CHRIS WALCEK

1 INTRODUCTION

Some trace gases and pollutants are readily removed from the atmosphere by becoming incorporated into cloudwater and then falling to Earth's surface in precipitation. In cloud-free air, small amounts of condensable species such as sulfates, ammonia, and nitrates coagulate with water vapor to form or nucleate extremely small atmospheric aerosol particles. Through gaseous diffusion and aerosol coagulation, smaller aerosols generally grow in size with time, while continuously maintaining an approximate equilibrium with water vapor and other surrounding condensable trace gases. Some aerosols eventually become large enough to develop an appreciable fall speed that overcomes frictional air drag, and thus slowly settle toward Earth's surface. However, well before particles grow to sizes where gravitational settling becomes important, larger aerosol particles are readily incorporated into clouds when aerosol-laden air cools during lifting, mixing with colder air, or other radiative cooling processes. Cloud drops initially form directly on aerosol particles, immediately incorporating a significant fraction of aerosol-borne trace constituents into the liquid phase. Once clouds form, soluble gases rapidly diffuse toward and dissolve into cloud droplets, contributing to trace chemical concentrations in cloudwater.

Most of the time, cloudwater evaporates, releasing the aerosols and gases that were absorbed during condensation back into the atmosphere. Under some conditions, cloud drops and ice particles grow by vapor diffusion and coalesce with other cloud drops and become large enough to develop appreciable fall speeds, at which point precipitation-sized cloud particles are formed. Within a cloud, falling liquid

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

and ice precipitation rapidly scavenges and accretes smaller cloud drops, and precipitation particles grow large enough to leave the cloud and fall to the surface. Outside the cloud, falling precipitation evaporates before reaching the surface, which releases some dissolved constituents back to the atmosphere, and concentrates the remaining dissolved constituents in precipitation before reaching the surface.

Thus there are many pathways through which trace constituents are transferred from the gas phase into precipitation at Earth's surface. Within a cloud there is nucleation scavenging, which takes place as ice and liquid water condenses on condensation and ice nuclei. Impaction scavenging of aerosols and small cloud particles by other cloud drops and other classes of hydrometeors (cloud ice, graupel, snow etc.) involves collisions between hydrometeors and interstitial cloud particles. Gases are also readily absorbed by all categories of hydrometeors through direct gaseous diffusion. Below cloud base, falling precipitation absorbs gases through diffusion, and aerosols are incorporated into falling precipitation by impaction scavenging.

Rudimentary calculations of the physics of these scavenging processes, as well as more sophisticated simulations reveal that the largest fraction of trace constituents in precipitation usually originate from the direct nucleation scavenging of soluble aerosols by cloud drops when they initially condense during cloud formation. Probably the next most important source of trace constituents in precipitation arises from the dissolution of soluble trace gases into cloudwater. Other scavenging mechanisms, such as impaction scavenging of aerosols by cloud or precipitation drops, and diffusion or impaction of gases or aerosols by larger precipitation particles are typically much less efficient. In the following sections, these important scavenging pathways are further discussed and quantified.

2 NUCLEATION SCAVENGING

It is well established that in the atmosphere the phase transition from water vapor to liquid water depends on the presence of cloud condensation nuclei (CCN). CCN are composed of water-soluble substances that bind with water molecules and significantly lower the equilibrium partial pressure of water vapor, allowing water to condense or change phase into these "salty" solutions when water vapor concentrations are well below saturation with respect to pure water.

Raoult's Law. The equilibrium partial pressure of water vapor over a solution containing a dissolved salt is very close to the mole fraction of water molecules in the solution times the saturated partial pressure of water vapor over pure water. This reduction of the equilibrium vapor pressure of water over a "salty" solution is known as "Raoult's law." Thus a droplet composed of 50% water molecules and 50% sodium chloride ions (25% Na^+ and 25% Cl^-) will be at equilibrium with an environment where the relative humidity is 50%. Therefore, the mole fraction of water molecules in "wetted" aerosols is very close to the ambient relative humidity, and therefore at humidities close to 100%, CCN and the wetted aerosols become

nearly "pure" water solutions. Even in the cleanest environments, there are several tens to hundreds of CCN per cubic centimeter of air. Therefore, there are always small amounts of wetted surfaces present in the atmosphere, and the time it takes H_2O vapor to diffuse toward these wetted aerosols is sufficiently small so that condensation or evaporation occurs rapidly, maintaining the environment near a saturated equilibrium with respect to these solution droplets at all times.

Clouds form when relative humidities (RH) exceed 100%, and, since the equilibrium partial pressure of water vapor is only a function of temperature, cloud formation is usually induced by the cooling of air. If the total amount of water in an air parcel remains the same, as an air parcel cools, the relative humidity rises, and a small amount of water condenses onto the wetted aerosols, and they swell in size. When air is cooled at humidities below 100%, the wetted aerosol absorbs water vapor, increasing the liquid-phase mole fraction, thus changing the equilibrium pressure of water vapor over the wetted aerosols following Raoult's law. However, when the RH exceeds 100%, the aerosol solutions become essentially "pure water," and additional water added to the droplets does not increase the equilibrium partial pressure of water vapor. Therefore, all vapor in excess of saturation rapidly condenses. Models and measurements in clouds show that the RH in clouds rarely exceeds about 101%, and typical supersaturations in a cloud are on the order of a few tenths of a percent (RH = 100.1 to 100.5%).

Kelvin Effect. For extremely small spherical drops, there is insufficient surface tension to "hold" condensed water in a liquid phase, and thus the equilibrium partial pressure of water vapor over a spherical droplet is higher than the equilibrium partial pressure over a flat liquid surface. For example, a spherical liquid drop with a radius of 0.01 μm requires a relative humidity of about 110% to maintain its size without evaporating. A pure water drop with a radius of 1 μm requires RH = 100.1% to maintain its size without growth or evaporation.

Köhler Curves. The *increase* of the partial pressure over a spherical surface due to the Kelvin effect counters the *reduction* in vapor pressure due to Raoult's effect. Together, the Raoult and Kelvin effects produce the classical "Köhler curve" (Köhler, 1926) describing the equilibrium partial pressure (e_{CCN}) over a spherical droplet of radius r containing a specified amount of dissolved ions relative to pure water:

$$\frac{e_{CCN}}{e_s} \approx 1 + \frac{a}{r} - \frac{b}{r^3} \tag{1}$$

where e_s is the equilibrium partial pressure of water vapor over a flat pure water surface, and the term involving a/r is a curvature (Kelvin) term, and the b/r^3 term is the solution (Raoult) term. Numerically, $a \approx 3.3 \times 10^{-5}/\mathrm{T}$ (K) (cm), and $b \approx 4.3\, im_s/M_s$ (cm^3) where i is approximately the number of dissolved molecules produced when the soluble CCN dissolves [e. g., sodium chloride (NaCl) produces 2 ions; sulfuric acid (H_2SO_4) produces 3], M_s is the molecular weight and m_s is the

dry mass of the soluble component of the CCN ($= \frac{4}{3}\pi r_s^3 \rho_s$, r_s = dry salt particle radius, ρ_s = salt density).

The equilibrium vapor pressure over a solution droplet may be larger or smaller than the vapor pressure over a plane surface of pure water, depending on whether the solute term is smaller or larger than the curvature term. Figure 1 shows examples of the saturation vapor pressure around drops that have condensed on two sizes and typical compositions of soluble aerosols. These curves show that there is a local maximum that occurs at a *critical* radius $[r_c = (3b/a)^{1/2}]$ and critical supersaturation $[s_c = 100(e/e_s - 1) = (4a^3b/27)^{1/2}]$. If a cloud drop is smaller than the critical size, and the local water vapor pressure is less than the critical supersaturation, drops grow only until they reach equilibrium with respect to the environmental water vapor pressure. If the water vapor concentration is greater than the critical supersaturation, a CCN will grow indefinitely, and a cloud drop is considered "activated" or "nucleated." Notice that the critical supersaturation is a strong function of the dry aerosol radius, with smaller particles requiring a higher supersaturation before they are activated relative to larger dry aerosols.

Within a cloud updraft, the saturation vapor pressure is continuously decreasing as air adiabatically expands and cools during lifting. When the RH initially exceeds

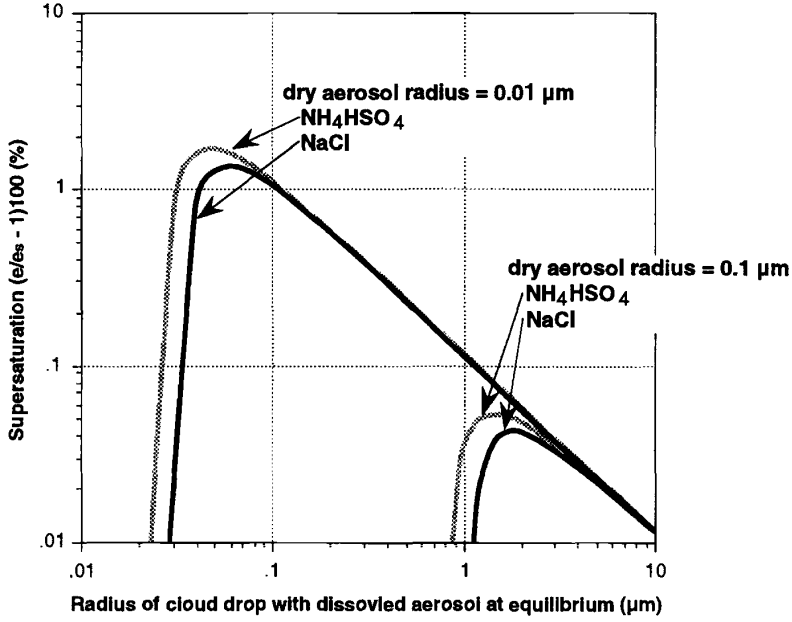


Figure 1 Equilibrium partial pressure of water vapor (expressed as supersaturation percent) as a function of the radius of a droplet containing a specified amount of dissolved salt (specified in terms of dissolved dry aerosol radius).

100%, water vapor primarily diffuses toward the few CCN containing the largest amounts of dissolved salts, which usually are the largest dry aerosols with the lowest critical supersaturations. Once a few cloud drops are activated, "excess" water vapor can either condense on existing drops or water vapor can diffuse toward and activate new CCN, depending on how rapidly the parcel is cooling. If not enough droplets are activated, the supersaturation increases, and more CCN are activated, incorporating smaller dry soluble aerosols into the cloud water. If sufficient numbers of nucleated drops present, water vapor diffuses toward existing drops, the supersaturation does not increase, and no additional drops are "activated."

The total number of cloud droplets nucleated during cloud formation depends in a very complicated way on the rate of cooling (i.e., lifting rate, or updraft velocity), and the size spectra and composition of soluble aerosols present in the cloud updraft.

Figure 2 shows one example of an explicit simulation of water vapor diffusion in an aerosol-laden cloud updraft. Supersaturations in a cloud updraft increase during the initial few meters above cloud base, and the greatest supersaturations are reached within 20 m or less above cloud base, usually within a few seconds above cloud base. Above the level of highest supersaturation, no additional CCN are activated, and condensing water vapor readily diffuses toward the already activated and growing cloud drops, which provide adequate surface area for condensation.

Numerous measurements show that condensation nuclei concentrations are typically a factor of 3 to 10 higher in continental areas relative to maritime areas, depending on the supersaturation at which CCN concentrations are measured. Figure 2 shows that in maritime environments containing relatively few CCN, supersaturations in a cloud updraft reach considerably higher values since there are so few

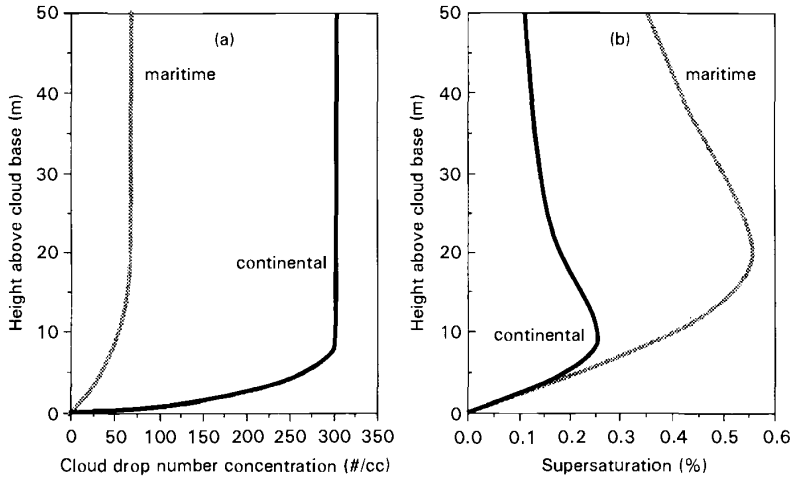


Figure 2 Initial development of cloud properties in air ascending at 1 m/s (*a*) number of cloud drops and (*b*) supersaturation. Typical maritime and continental CCN spectra assumed.

condensation nuclei present during cloud formation. Therefore, fewer drops are ultimately nucleated. In contrast, in continental areas where significantly greater numbers of CCN are present, more cloud drops form, and since vapor more efficiently diffuses toward these drops, supersaturations are appreciably lower than maritime clouds.

Figure 3 shows typical droplet number concentrations, maximum supersaturations, and the minimum dry radii of aerosols nucleated in a cloud updraft according to the Köhler theory following the approach of Twomey (1959). The influence of different aerosol size distributions on cloud properties is crudely accounted for here by assigning typical "continental" and "maritime" CCN distributions. Usually maritime conditions have lower numbers of CCN, and the differences between the maritime and continental distributions shown in Figures 2 and 3 qualitatively show the range of the natural variations that occurs in various cloud environments. This figure shows that greater numbers of cloud drops are nucleated at higher updraft velocities, due to increased cooling rate, and therefore condensation rate within the rising cloud parcel. Under continental conditions containing greater numbers of CCN and aerosols, more drops are nucleated and peak supersaturations are lower.

Figure 4 shows the fraction of soluble aerosol mass that is activated or incorporated into cloud drops during cloud formation for several updraft velocities under typical "maritime" or "continental" CCN and aerosol distributions, estimated two ways. One method is to add up the mass of the largest aerosols observed in a typical

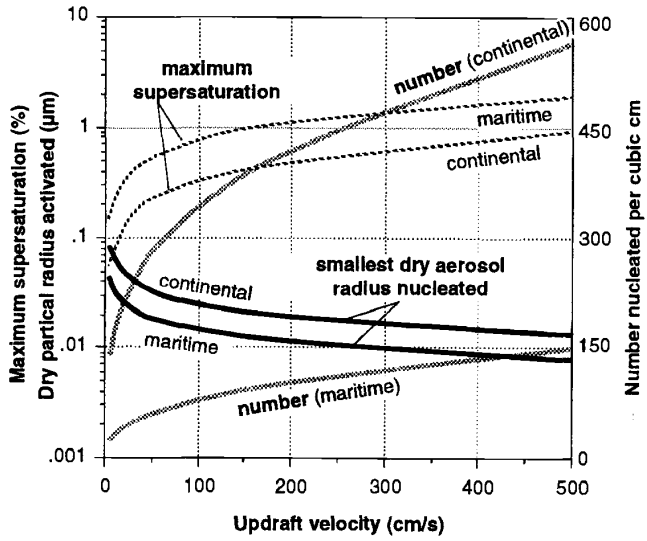


Figure 3 Maximum supersaturation, dry radius of the smallest aerosol activated, and number of cloud drops formed during condensation within a cloud updraft. Ammonium bisulfate aerosol, and typical supersaturation activation for maritime or continental air masses assumed.

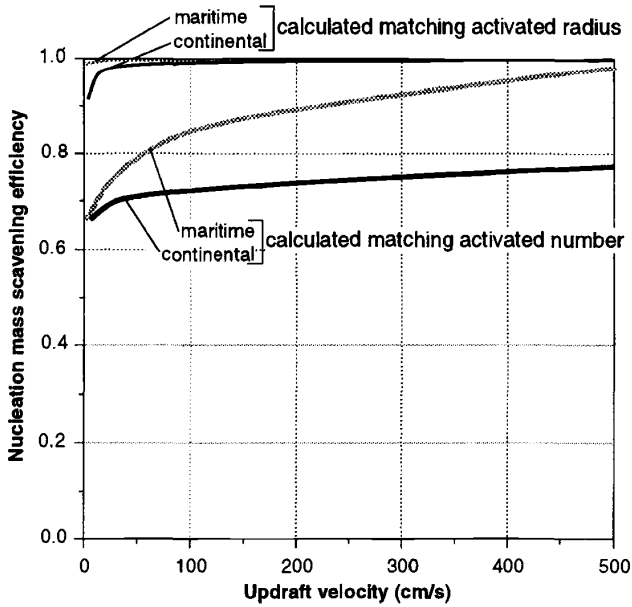


Figure 4 Fraction of aerosol mass scavenged during condensation within a cloud updraft. Activation spectra and dry aerosol size distributions typical of continental or maritime for ammonium bisulfate aerosol assumed.

dry aerosol size distribution that were nucleated during condensation. Thus if 300 cloud drops are nucleated, one can calculate the mass of the 300 largest dry aerosols in a measured aerosol size distribution, and compare this mass to the total aerosol mass in all sizes. Another method for estimating nucleation scavenging involves using the Köhler equation. Knowing the maximum supersaturation in a cloud updraft, one can calculate the size of the smallest soluble aerosol particle nucleated. Knowing this size, one can calculate the mass fraction contained in aerosol particles greater than this size from a measured aerosol size distribution. These two methods yield slightly different mass fractions and suggest that there are some inconsistencies and uncertainties in our scientific understanding of the nucleation processes and how it relates to the size distribution of dry aerosols entering a cloud. Despite these uncertainties, Figure 4 shows that a large fraction of the mass of soluble aerosols is activated and incorporated into cloudwater when a cloud forms. Irrespective of these minor uncertainties, the fraction of aerosol mass nucleated is proportional to the updraft velocity and inversely related to the number concentration of aerosol in air. Thus clouds forming under continental conditions typically scavenge a slightly smaller fraction of the aerosol mass.

The concentration of aerosol-laden trace constituents in cloud water can be given by:

$$C_l = \frac{\varepsilon_{\text{aer}} 10^6 C_T}{L} \tag{2}$$

where C_l is the liquid-phase concentration (moles per $\text{liter}_{\text{water}}$), ε_{aer} is the mass scavenging fraction of the aerosol-borne trace constituent, shown in Figure 4, C_T is the total concentration (moles per $\text{liter}_{\text{air}}$) of trace constituent in air from which the cloud forms, and L is the condensed water content of the cloud (grams water per cubic meter of air). As shown in Figure 4, the mass-scavenging fraction is usually large, and nearly all aerosol-borne constituents are incorporated into the aqueous phase within clouds during condensation. Low scavenging efficiencies for soluble aerosols occur under highly polluted, high particle number concentration conditions, or within clouds that form slowly at low cooling rates, such as fogs.

Trace Gas Scavenging

After clouds form, soluble gases rapidly diffuse toward and dissolve into the liquid phase. In the presence of liquid water, gases partition themselves between gas and aqueous phases, and the liquid-phase concentration (moles per liter) divided by the partial pressure (atm) of the dissolved constituent over the liquid at equilibrium is defined as the Henry's law coefficient (K_h moles/liter/atm), a standard measure of trace gas solubility.

For typical clouds, interstitial gases diffuse toward and establish an equilibrium with a condensed phase within a few seconds or less. Therefore, soluble gases are very close to Henry's law equilibrium with cloud drops. Under equilibrium conditions, the liquid-phase concentration of a soluble gas in cloudy air can be written in terms similar to the expression for the concentration of soluble aerosol [Eq. (2)]:

$$C_l = \frac{\varepsilon_{\text{gas}} 10^6 C_T}{L} \tag{3}$$

where C_l is the liquid-phase concentration (moles per $\text{liter}_{\text{water}}$), and as previously defined for aerosols, C_T is the total concentration (moles per $\text{liter}_{\text{air}}$) of trace gas in the cloudy air. The "scavenging efficiency" (ε_{gas}) of soluble gases can be calculated from mass conservation and equilibrium constraints as

$$\varepsilon_{\text{gas}} = \frac{1}{10^6/(K_h LRT) + 1} \tag{4}$$

Here R is the universal gas-law constant (0.082 atm liter/mol K) and T is the temperature (K). For highly soluble gases that partition predominantly into the liquid phase, the term involving K_h in $(4) \ll 1$, ε_{gas} is close to unity, and $C_l = 10^6 C_T/L$. At the other extreme, for low-solubility gases that remain predominantly in the gas phase, the K_h term in $(4) \gg 1$, ε_{gas} is small, and therefore $C_l = C_T K_h RT$, independent of the cloud liquid water content.

Many gases rapidly dissociate into several chemical forms when they dissolve in cloud water. For example, SO_2 is a weak acid dissociating into three chemical forms: a nonionic hydrated complex (SO_2H_2O), bisulfite (HSO_3^-), and sulfite ($SO_3^=$) ions. Equilibrium expressions are defined to quantify this dissociation as

$SO_2(g) \leftrightarrow SO_2H_2O$ (aq)	$K_h = [SO_2H_2O(aq)]/P_{SO_2}(g)$ mol/liter atm
$SO_2H_2O(aq) \leftrightarrow HSO_3^- + H^+$	$K_1 = [HSO_3^-][H^+]/[SO_2H_2O$ (aq)] mol/liter
$HSO_3^- \leftrightarrow SO_3^= + H^+$	$K_2 = [SO_3^=][H^+]/[HSO_3^-]$ mol/liter

In addition to Henry's law constant for SO_2, laboratory-measured equilibrium coefficients K_1 and K_2 are used to quantify the first and second dissociation of SO_2 in solution. Other gases such as organic acids, CO_2, NH_3, HNO_3, and HCl dissociate in a similar manner, and one can define an "effective" Henry's law solubility, which accounts for the concentrations of *all* dissolved chemical forms of the trace gas, which can generally be expressed as:

$$K_{he} = K_h\left[1 + \frac{K_1}{[H^+]}\left(1 + \frac{K_2}{[H^+]}\right)\right] \tag{5}$$

This effective solubility must be used in (4) for estimating liquid-phase concentrations in a cloud. This effective Henry's law solubility is therefore usually a function of the concentration of hydrogen ion $[H^+]$, or the acidity of the cloudwater, which is proportional to the concentrations of the acids and bases in cloudy air, and strongly influenced by the cloud liquid water content.

Figure 5 shows the mass-scavenging fraction for gases as a function of their effective Henry's law solubility. Also shown on this figure is the approximate solubility and scavenging fraction of several trace gases of interest in atmospheric chemistry. For gases with Henry's law constants less than about 100 mol/liter atm, only an extremely small fraction enters into the liquid phase in a typical cloud. This includes most organic gases, NO, NO_2, and CO, on the order of a few percent of formaldehyde and SO_2 dissolve in a cloud. In contrast, 50 to 80% of hydrogen peroxide (H_2O_2) dissolves in most clouds, and nearly all ammonia (NH_3), nitric acid (HNO_3), and sulfuric acid (H_2SO_4) are scavenged within cloudy air.

Wet Deposition Fluxes

The rate at which trace chemicals are removed from the atmosphere via wet deposition is intimately related to the life cycle of liquid water in the atmosphere. The flux of dissolved constituents to the surface in precipitation is the product of the precipitation rate P_r (mm/h = kg_{water}/m^{-2} h) and the liquid-phase concentration of trace constituents in precipitation (C_l moles per liter of solution)

$$\text{Flux (mol/m}^{-2}\text{/h)} = C_l P_r \tag{6}$$

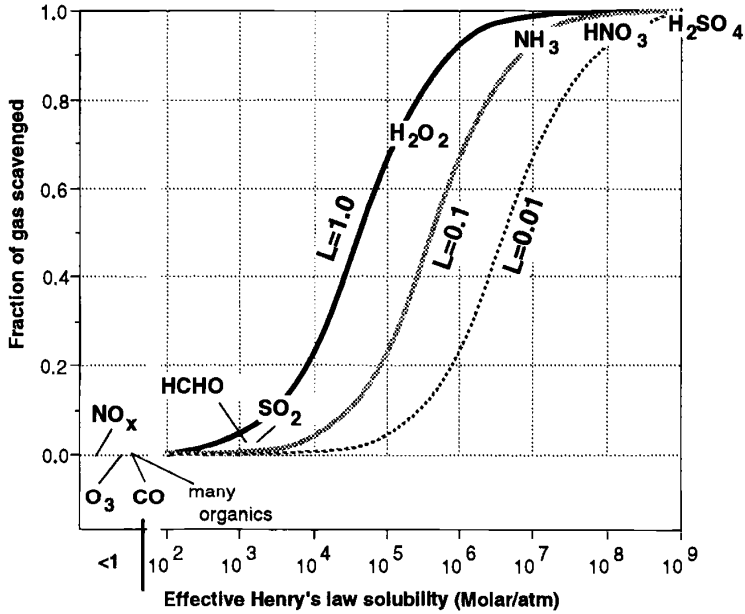


Figure 5 Fraction of gases scavenged in the presence of cloud liquid water as a function of the effective Henry's law solubility. Scavenging fraction shown for several cloud liquid water contents (L, grams liquid water per cubic meter). Approximate solubility of numerous gases shown.

Since precipitation forms by collecting and scavenging small cloud droplets from throughout a cloudy layer, the concentration of dissolved constituents in precipitation to a first approximation is proportional to the average concentration of constituents in the cloudwater from which the precipitation forms. Evaporation of precipitation below cloud base increases concentrations in precipitation, especially during the initial stages of a precipitation event. As noted above, the concentrations of aerosols and highly soluble gases in cloudwater are inversely proportional to the condensed water content in a cloud, which is in turn proportional to the amount and speed of cooling during cloud formation. The amount of cooling is proportional to the amount of lifting in convective or orographic clouds, and other radiative factors determine the cooling rate in fogs and some stratiform clouds. Thus the water content of a cloud is often proportional to the depth or vertical displacement of air within a cloud, although entrainment and mixing of dry air from outside a cloud often evaporates and dilutes condensed water in clouds.

Liquid water contents in clouds typically increases with altitude above cloud base, and therefore liquid-phase concentrations generally decrease with altitude above cloud base as the dissolved gases and aerosols are diluted by the increasing amounts of liquid water.

Figure 6 shows vertical profiles of "adiabatic" and "typical" water contents within a convective cloud updraft. Here, adiabatic water contents are calculated assuming that the updraft remains saturated as it cools, and *all* water vapor in excess of saturation remains in the updraft as liquid water. Adiabatic water contents are rarely observed in clouds and represents an upper limit to the amount of condensed water within a cloud. Measurements (Warner, 1970) typically show that water contents in convective clouds are ~20 to 40% of adiabatic values.

Since liquid-phase concentrations are proportional to total concentrations of trace substances in air [Eqs. (2)–(4)], one can derive a simple expression for the rate of change of a trace chemical concentration in a precipitating environment due to precipitation scavenging:

$$\frac{dC_T}{dt} = -\frac{\varepsilon 10^3 P_r}{\bar{L}\,\Delta z} C_T = -\frac{C_T}{\tau_s} = -\left(\frac{\varepsilon}{\tau_{cw}}\right) C_T \tag{7}$$

where Δz is the depth of the cloudy layer experiencing precipitation, and L is the water content averaged over the cloud depth (g/m^3); τ_s is a time constant for soluble species to be removed from the cloudy environment due to precipitation scavenging,

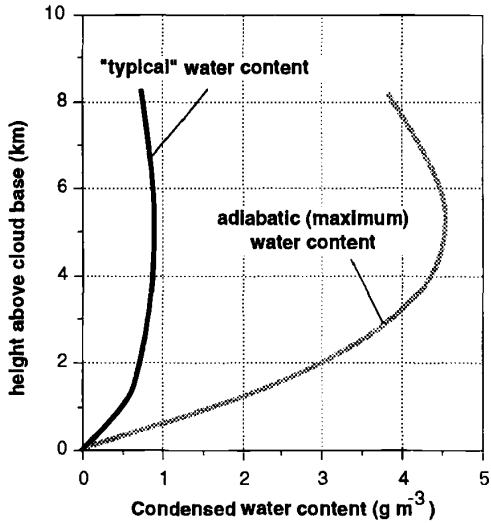


Figure 6 Condensed water content within a convective updraft vs. height above cloud base. Upper limit (adiabatic) and typical horizontal average through a cloud shown. Cloud base 10°C at 900 mbar. Typical water contents are adiabatic water contents scaled by Warner's (1970) compilation of observed reduction factors.

and is proportional to the time constant for the removal of condensed water τ_{cw} in a cloud; ε is the scavenging efficiency for either soluble aerosols (Fig. 4) or trace gases [Eq. (4), Fig. 5]. The time constant for the removal of condensed water in a cloud is the condensed water path in a cloud (mm $= L\,\Delta z/10^3$) divided by the precipitation rate (mm/h) from the cloud. For highly soluble species, the scavenging efficiency shown in Figures 4 and 5 are very high ($\varepsilon \sim 1$), and therefore the wet deposition time scale for liquid water is identical to the time constant for removal of soluble species from the cloudy environment.

Using reasonable estimates for the parameters in Eq. (7), one can quantify the wet deposition time scales for the removal of trace constituents under precipitating conditions. For a convective cloud, Figure 7 shows precipitation rate and the time scale for washout of condensed water as a function of cloud depth at various storm efficiencies. Updraft at cloud base is 1 m/s, and precipitation rates (and washout times) scale linearly with this assumed velocity. Storm efficiency is defined here as the surface precipitation rate divided by the condensation rate in the updraft. Storm efficiencies range from 10%, in relatively dry, high-wind-shear environments to 100% in saturated, low-shear environments (Weisman and Klemp, 1982; Lipps and Hemler, 1986).

Average condensed water contents for calculating the washout lifetime are taken from "typical" water contents shown in Figure 6, or closer to adiabatic (shaded gray region on Fig. 7*b*) conditions. The main point of Figure 7 is to show that time scales for removal of condensed water substance are on the order of an hour or less, and therefore soluble constituents that are completely absorbed into cloudwater are removed rapidly from the atmosphere when precipitation is occurring. For less soluble constituents, such as SO_2, which partitions only a few percent into the aqueous phase in a cloud, the time constant for wet removal is longer by a factor of 10 to 100, depending on the cloud depth and microphysical storm efficiencies.

Therefore we conclude from these semiqualitative estimates that in the immediate vicinity of precipitation systems, soluble gases and a large fraction of the mass of soluble aerosols are efficiently removed from the atmosphere on a time scale of an hour or less. Chemical species that are rapidly removed by precipitation include gaseous HNO_3, NH_3, and H_2O_2. Soluble constituents of CCN including aerosol sulfates, nitrates, and sea salts are also rapidly scavenged from the atmosphere under most precipitating conditions.

From a global perspective, the rate-limiting factor determining how quickly trace constituents are removed from the atmosphere is determined by how often and where precipitation occurs. At any given time, clouds typically cover approximately half of Earth's surface, but only a small fraction of clouds are precipitating. Pruppacher and Klett (1978) suggest that from a global perspective the average residence time of condensed water in the atmosphere is on the order of 7 h, and the residence time of all water substance (vapor + condensed water) is on the order of 9 days. Therefore on a global perspective we expect similar removal time scales for soluble trace chemicals, scaled by the relative partitioning of trace gases in condensed and vapor phases.

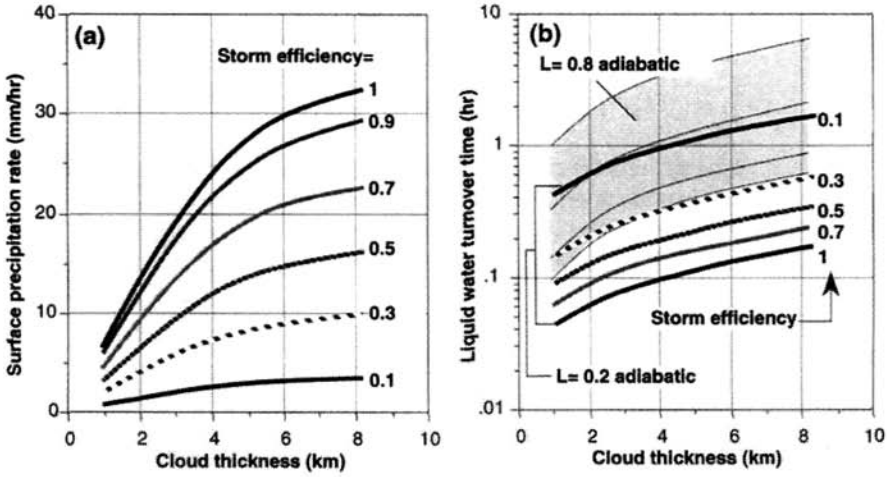


Figure 7 (*a*) Precipitation rate and (*b*) condensed water washout time scale within convective clouds as a function of cloud depth at various storm efficiencies. Updraft at cloud base is 1 m/s, and precipitation rates (and washout times) scale linearly with this assumed velocity. Storm efficiency defined as the surface precipitation rate divided by the condensation rate in the updraft. Average condensed water contents for calculating lifetime taken from typical water contents shown in Figure 6 or closer to adiabatic (shaded gray region).

Acid Deposition

An "acid" is essentially any substance that releases hydrogen ions (H^+) when dissolved in water. Several atmospheric trace constituents dissociate into positive and negative ions when dissolved in water, and some are acidic to varying degrees. The strongest acids in atmospheric waters are dissolved sulfuric acid (H_2SO_4) and nitric acid (HNO_3). Numerous other acidic substances have been identified in the atmosphere, such as sulfur dioxide (SO_2), organic acids, hydrochloric acid (HCl), carbon dioxide (CO_2), and even water itself, but these substances are either relatively weak acids or are present at relatively small concentrations and thus do not usually contribute appreciably to measured acidity. For any solution, there is an equal concentration of dissolved positive and negative ion charge, and a typical ion balance in cloudwater and precipitation is

$$[H^+] + [Na^+] + [NH_4^+] + [\text{soil ions}] = (\text{positive ions})$$
$$= 2[SO_4^=] + [NO_3^-] + [Cl^-] + [HCO_3^-] \ (\text{negative ions})$$

Na^+ and Cl^- ions arise from dissolved sea salt aerosols and are usually present in approximately equal concentrations. NH_4^+ is dissolved ammonia, "soil ions" refers to calcium and magnesium cations that are typically associated with carbonates

(HCO_3^-) in soil dust, and $SO_4^=$ and HNO_3^- are sulfuric and nitric acid. Since an ion balance is always maintained in atmospheric waters, the concentration of H^+ is

$$[H^+] = 2[SO_4^=] + [NO_3^-] + [HCO_3^-] - [NH_4^+] - [\text{soil ions}]$$

Thus the concentration of the hydrogen ion is proportional to the concentrations of sulfates, nitrates, bicarbonate (dissolved CO_2), ammonia, and carbonate-laden soil dust dissolved in cloudwater and precipitation. Measured concentrations of H^+ vary over several orders of magnitude, and therefore a logarithmic pH scale is used to quantify acidity levels in water

$$pH = -\log_{10}[H^+]$$

Using the pH scale, a decrease in one pH unit corresponds to a 10-fold increase in acidity or H^+ concentration. Also, as the pH decreases, the concentration of H^+ and acidity increases. Pure water in equilibrium with atmospheric CO_2 has a pH near 5.6, but the concentrations of sulfate, nitrates, ammonia, or soil cations in cloud and rainwater usually greatly exceed the concentrations of dissolved CO_2, even in remote areas. Typical "clean" atmospheric waters have a pH of 4.5 to 5.5. In more polluted areas, pHs in precipitation range from 3 to 4, and in some low liquid water content clouds, pHs as low as 2 to 3 have been measured.

Formation of Acids. Sulfuric and nitric acids are produced by reactions between atmospheric oxidants and emitted sulfur and nitrogen oxides (SO_2 and NO_x), which are by-products of fossil fuel combustion and other industrial activities. The dominant reactions converting SO_2 to sulfuric acid include reactions with hydrogen peroxide (H_2O_2) in clouds and hydroxyl radical (HO) in air. Nitric acid is produced by the oxidation of NO_2 by the HO radical, and also at night by a heterogeneous reaction involving ozone, NO_2, and NO_3 radicals. Atmospheric oxidants responsible for acid formation are produced via a complex sequence of photochemical reactions, and some acid-generating chemical reactions occur among dissolved gaseous constituents in atmospheric clouds or aerosols. Typically, emitted NO_x is converted to nitric acid within a day or less, and SO_2 is converted to sulfuric acid within several days following emission. The concentrations of the oxidants, and the time scale for chemical reactions vary strongly with season, latitude, time of day, sunlight intensity, background concentrations of NO_x and organic compounds, and many other chemical and meteorological factors.

Strong acids have an affinity for water, and therefore hygroscopically grow or combine with water vapor to form "haze" aerosols containing sulfuric acid, nitric acid, and varying degrees of neutralizing ammonia (NH_3), especially when atmospheric relative humidities are above 60 to 70%. Typically, ammonia and nitric acid are present as both gases and aerosols in the atmosphere, while sulfate partitions predominantly into condensed aerosols. These sulfate, nitrate, and ammonium-containing aerosol particles constitute a significant fraction of cloud condensation nuclei (CCN), and thus acid-containing aerosols are readily incorporated into clouds.

Precipitation forming within clouds therefore contains dissolved CCN together with other soluble gases such as HNO_3 and NH_3.

Undesirable Effects. At high concentrations or exposures, acidic solutions induce numerous undesirable reactions with surfaces. In conjunction with other pollutants, acid deposition contributes to potentially deleterious effects on aquatic, agricultural, and forest ecosystems. Chemical changes attributed to the deposition of acidity from the atmosphere have been measured in forest ecosystems and surface waters.

Concentrations of acids in lakes have been correlated with the concentrations and deposition rates of atmospheric acids, and high concentrations of acids in lakes and streams can adversely affect fish populations. Health effects associated with exposure to acid-containing particulates in humans remains an area of uncertainty since current studies of these effects are too limited to unambiguously discern dose–response relationships in humans. Acid deposition from the atmosphere has been shown to accelerate the deterioration rate of exposed metals, painted finishes, and concrete or stone surfaces.

In industrialized areas, concentrations of sulfuric and nitric acids in cloud water and precipitation are up to 50 to 100 times greater than values measured in areas that are not influenced by upwind emissions of anthropogenic pollutants. The relative concentrations of deposited sulfur and nitrogen acids are correlated with the relative emission rates of sulfur and nitrogen pollutants over larger areas.

The amount of acidity within precipitation is strongly influenced by numerous meteorological and chemical factors, as well as the emission rates of precursor sulfur and nitrogen pollutants. Therefore a thorough understanding of larger-scale meteorology, cloud dynamics and microphysics, and atmospheric chemistry is required to fully quantify and study atmospheric acidity

REFERENCES

Köhler H., Zur thermodynamic der kondensation an hygroskopichen kernen und bemerkungen über das zusammenfliessen der tropfen, *Medd. Met. Hydr. Anst. Stockholm*, 3(8), 1926.

Lipps, F. B., and R. S. Hemler, Numerical simulation of deep tropical convection associated with large-scale convergence, *J. Atmos. Sci.*, *43*, 1796–1816, 1986.

Pruppacher, H. R., and J. D. Klett, Microphysics of Clouds and Precipitation, D. Reidel Publishing, 714 pp., 1978.

Twomey, S., The nuclei of natural cloud formation: The supersaturation in natural clouds and the variation of cloud droplet concentration, *Geophys. Pura Appl.*, *43*, 243–249, 1959.

Warner, J., *J. Atmos. Sci.*, *27*, 1035, 1970.

Weisman, M. L., and J. B. Klemp, The dependence of numerically simulated convective storms on vertical wind shear and buoyancy, *Monthly Weather Rev.*, *110*, 504–520, 1982.

CHAPTER 20

LARGE-SCALE CIRCULATION OF THE STRATOSPHERE

WILLIAM L. GROSE

1 GOVERNING EQUATIONS

In their most basic form, the governing equations that determine the circulation and thermal structure of the atmosphere are referred to as the primitive equations. The following version of those equations do include some approximations, and the reader is referred to a standard text such as Holton (1992) for a fuller discussion of the details and the advantages of using pressure, p, as a vertical coordinate. In vector form the horizontal momentum equation can then be written in an (x, y, p) coordinate system as

$$\frac{DV}{Dt} = -f\mathbf{k}xV_h - \nabla_p\Phi \tag{1}$$

with the total derivative given as

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + \omega\frac{\partial}{\partial p}$$

and ∇_p is the horizontal gradient operator (with partial derivatives taken holding p constant). Here V_h is the horizontal velocity, Φ is the geopotential, f is the Coriolis parameter, and $\mathbf{k}$ is a unit vector in the vertical direction. Also, x is the west-to-east coordinate, y is the south-to-north coordinate, and u and v are the zonal and meri-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

dional components, respectively, of the horizontal velocity, V_h. The variable, ω is defined as

$$\omega = \frac{Dp}{Dt}$$

and becomes the surrogate for the vertical velocity, w, in the (x, y, p) coordinate system.

For large-scale motions, the vertical component of velocity is typically several orders of magnitude smaller than for the horizontal velocity, and vertical accelerations can be neglected to a good approximation. The vertical component of the momentum equation then reduces to a diagnostic equation and can be expressed as

$$\frac{\partial \Phi}{\partial p} = \frac{RT}{p} \tag{2}$$

with R the gas constant for air and T the temperature.

The mass continuity equation in this coordinate system becomes

$$\nabla_p \cdot V_h + \frac{\partial \omega}{\partial p} = 0 \tag{3}$$

Finally, the thermodynamic energy equation becomes

$$\frac{DT}{Dt} = \frac{\omega RT}{C_p p} + \frac{Q}{C_p} \tag{4}$$

with C_p the specific heat at constant pressure for air. The variable Q is the diabatic heating rate. For the stratosphere, Q is typically the radiative heating rate and can be calculated as a function of the other dependent variables, knowing the distribution of radiatvely active species such as ozone, water vapor, and carbon dioxide (Goody, 1995).

Equations (1) to (4) represent the primitive equations in the (x, y, p) coordinate frame and form a determinate system of equations for the variables u, v, ω, Φ and T. These equations are inherently nonlinear and must be integrated in time with suitable initial and boundary conditions. Direct solution of the primitive equations requires the use of sophisticated numerical techniques implemented on digital computers.

A relationship of fundamental importance can be derived from the primitive equations [see Pedlosky (1979) for the derivation] in terms of Ertel's potential vorticity, Π, namely

$$\frac{D\Pi}{Dt} = 0 \tag{5}$$

for frictionless, adiabatic conditions. Here, Π, is given by

$$\Pi = \frac{1}{\rho}(\zeta + f)\nabla\Theta \tag{6}$$

Note that in Eq. (6) the total derivative is now expressed in the Cartesian coordinate system (x, y, z) as

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + w\frac{\partial}{\partial z}$$

Also, ∇ in Eq. (6) is now the three-dimensional gradient operator in the (x, y, z) coordinate system with ρ the density, ζ the relative vorticity (curl of the velocity V), and Θ the potential temperature. For frictionless, adiabatic flow, this relationship requires that Π be conserved following the motion. Physically, potential vorticity is representative of the ratio of the rotation of a fluid vortex column to the depth of the column. Conservation of potential vorticity for a compressible fluid can be thought of as analogous to conservation of angular momentum for a solid body. This conservation property for Π represents a very powerful constraint on the motions, particularly in the lower stratosphere where Eq. (5) is a reasonable approximation for time scales up to about 10 days. Distributions of Π on an isentropic surface (i.e., a constant potential temperature surface) thus represent a conserved dynamical tracer and, as we shall later see, provide much useful insight into the nature of the transport.

2 VERTICAL TEMPERATURE STRUCTURE

A traditional means of separating Earth's atmosphere into regions with quite distinct characteristics originated from considerations of the vertical temperature structure. A reference temperature profile for middle latitudes is shown in Figure 1 as a function of the geometric height using data from the compilation, *U. S. Standard Atmosphere* (1976). The difference between the stratosphere and troposphere in terms of the vertical temperature profile is readily apparent. Note that in the troposphere, the temperature T, decreases with increasing height z, up to the level of the tropopause. In contrast, the temperature in the lower stratosphere is nearly constant and then increases with increasing height through the middle and upper stratosphere until the level of the stratopause is reached. It is this difference in the temperature lapse rate, Γ, where

$$\Gamma = -dT/dz \tag{7}$$

that results in a difference in the stability characteristics between these two regions of the atmosphere.

The atmosphere is said to be statically stable, if after an air parcel is adiabatically displaced in the vertical dimension, the net forces acting on the parcel tend to restore

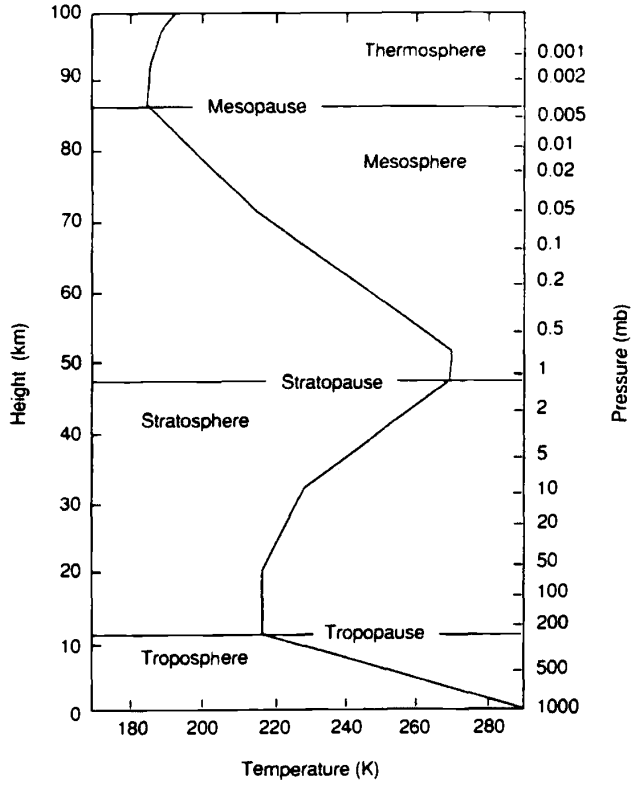


Figure 1 Mean temperature profile at midlatitudes based upon the *U. S. Standard Atmosphere* (1976) (from Holton, 1992).

it to its unperturbed position. The usual diagnostic for examining the static stability of the atmosphere is the buoyancy frequency or Brunt–Vaisala frequency (Holton, 1992), N, where

$$N^2 = [g(\Gamma_d - \Gamma)/T] \tag{8}$$

with g being the acceleration due to gravity and Γ_d being the dry adiabatic lapse rate (9.8°/km). The buoyancy frequency is the frequency of oscillation of an air parcel that has been adiabatically displaced from its equilibrium position in an atmosphere at rest. A situation with $N^2 > 0$ ($\Gamma < \Gamma_d$) corresponds to a stably stratified atmosphere. An inspection of Figure 1 reveals that N^2 would be greater than zero (Γ is zero or negative) in the stratosphere, and hence, this region should be stable. Indeed, in practice it has been found that vertical mixing in the stratosphere is much slower

than is typical for the troposphere, where air parcels can be transported vertically between the surface and the tropopause on time scales of several days or even as little as a few minutes in the case of very strong convective activity associated with the largest thunderstorms (Wallace and Hobbs, 1977).

This stable stratification with relatively slow vertical mixing is an important influence on the stratospheric circulation and gives rise to the long residence times in the stratosphere that have been deduced for long-lived constituents such as the chlorofluorocarbons (CFC) [appproximately 50 years for CFC-11 and 100 years for CFC-12, the two most abundant CFC compounds (WMO, 1994)]. Here, long-lived is used in the sense that the characteristic time scale for chemical change is very much greater than that associated with the transport of a constituent.

3 ZONAL-MEAN CLIMATOLOGY OF TEMPERATURES AND ZONAL WINDS

Climatologies of zonal-mean (spatial averages at constant latitude) values of temperatures and zonal (east–west component) winds are displayed as a function of height and latitude in Figure 2*a* and 2*b*, respectively. An inspection of the temperature cross section shown in Figure 2*a* reveals marked variations in temperatures between the winter and summer hemispheres in the stratosphere. Note that the tropopause above the equator occurs at a much higher altitude than above the polar regions and that distinct discontinuities or "breaks" occur in the tropopause at middle latitudes. In the lower, summer stratosphere the temperature increases from the equator toward the pole. In contrast, there is a midlatitude warm belt (Ramanathan and Grose, 1978) in the lower, winter stratosphere. Coldest temperatures occur in the lower stratosphere over the south polar region during winter. The northern polar regions exhibit more dynamical variability during winter and are typically not quite as cold. In the upper stratosphere the temperature increases monotonically from winter to summer pole.

The corresponding cross section of zonal winds is shown in Figure 2*b*. Note the presence of easterly winds (from east toward the west) in the summer hemisphere and westerly winds (from west toward the east) in the stratosphere. The seasonal reversal of summer easterlies to winter westerlies in each hemisphere is observed to be a prominent feature of the stratospheric circulation. The axis of the wintertime, westerly jetstream tilts poleward with decreasing height in the stratosphere resulting in a high latitude jetstream in the lower stratosphere, the so-called polar night jet.

4 ZONAL-MEAN MERIDIONAL CIRCULATION

The zonally averaged, meridional (north–south component) and vertical winds are, on average, at least an order of magnitude smaller than the zonal winds described in the previous section. A useful depiction of the circulation in the meridional plane (latitude vs. height), was originally derived by Murgatroyd and Singleton (1961).

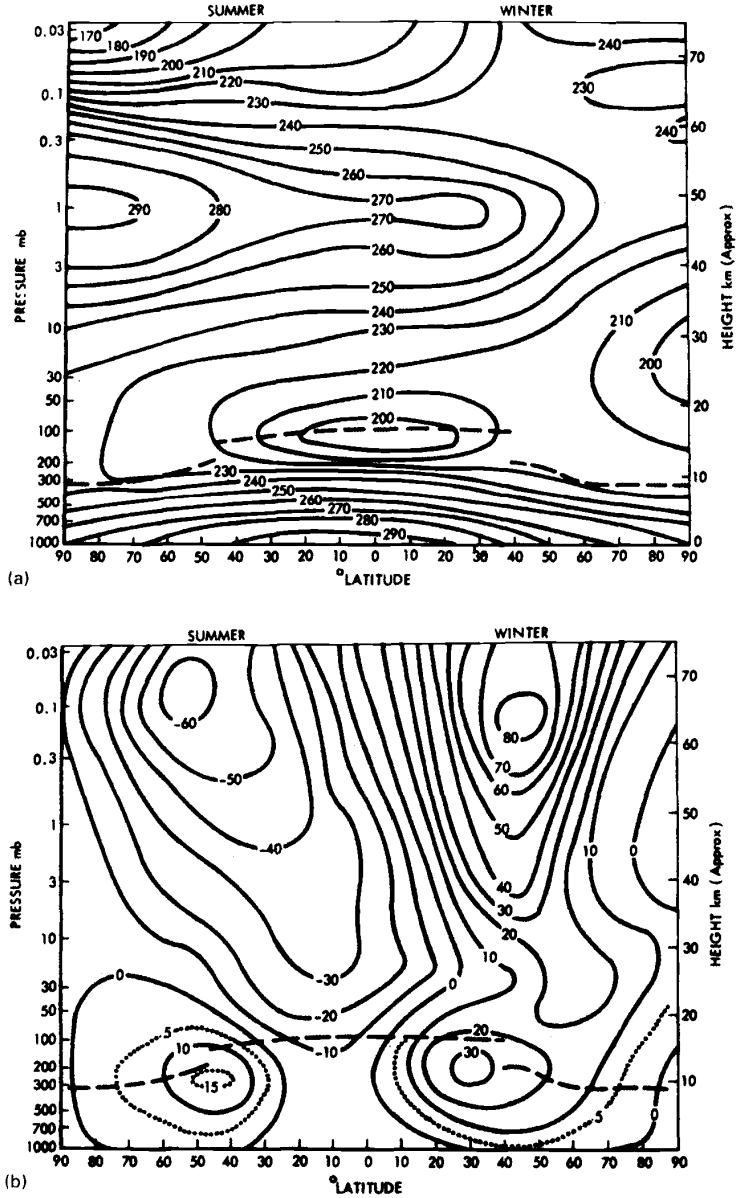


Figure 2 Latitude–altitude cross section at solstice conditions of zonally averaged (*a*) temperature (K) and (*b*) zonal wind component (m/s) (from Murgatroyd, 1969).

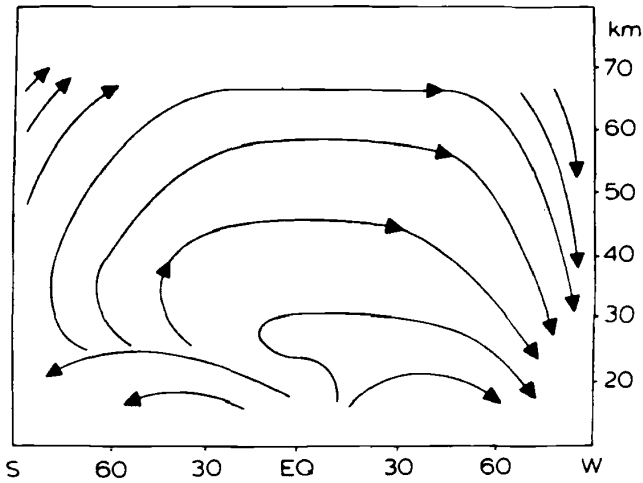


Figure 3 Schematic streamlines of the diabatic circulation for solstice conditions. S, the summer pole and W the winter pole (from Dunkerton, 1978).

This circulation is now generally referred to as the diabatic circulation. It is conceptually useful in that it illustrates the sense of the actual mass motions in the meridional plane. Using net radiative heating rates originally compiled by Murgatroyd and Singleton for solstice conditions, Dunkerton (1978) derived the vertical velocities necessary to balance these heating rates. From considerations of mass continuity, the corresponding meridional velocities were then calculated. Figure 3 shows the resultant streamlines inferred by Dunkerton (1978) from the calculated velocity fields. The large-scale stratospheric circulation in the meridional plane exhibits rising motion in the summer hemisphere with a slow (seasonal or longer time scale) meridional drift and subsidence over the winter pole (Andrews et al. 1987). A Coriolis torque associated with the meridional drift influences the production of the summertime (wintertime) zonal-mean easterlies (westerlies) in the stratosphere seen in Figure 2*b*.

5 WAVE MOTIONS

The depiction of the large-scale stratospheric circulation discussed in the preceding two sections is a zonally averaged perspective as noted. However, wavelike disturbances (commonly referred to as "waves"), which propagate vertically from the troposphere producing departures from zonal symmetry, are known to be important in determining the circulation and the transport of constituents in the stratosphere.

Important departures from the climatological zonal mean state shown in Figure 2*a* and 2*b* occur as a result of sudden stratospheric warmings, the quasibiennial oscillation (QBO) and the semiannual oscillation (SAO). Occurrence of these latter two phenomena is presently believed to result at least in part from vertically propagating Kelvin, Rossby-gravity, and/or gravity waves. The QBO manifests itself as an oscillation of the zonal winds with alternating westerlies and easterlies in the equatorial lower stratosphere. The period of the oscillation varies, but averages about 27 months. The SAO is also an oscillation of equatorial zonal winds with alternating westerlies and easterlies, but this phenomenon occurs in the upper stratosphere (and lower mesosphere) on a semiannual basis as the name implies. The mechanisms responsible for the QBO and SAO are quite different. The interested reader is referred to Andrews et al. (1987) for further details.

The sudden stratospheric warming phenomenon occurs during some years in association with anomalous enhancement of the amplitude of vertically propagating, planetary-scale disturbances into the wintertime stratosphere. Major warming events are characterized by very rapid increases in polar temperatures (50 to 70 K in a week or less) and severe disruptions of the wintertime westerly polar vortex with zonal easterlies replacing zonal westerlies at high latitudes. These warming events can occur in either hemisphere, although generally Southern Hemisphere warming events tend to be somewhat less spectacular than their counterpart in the Northern Hemisphere.

The various different types of waves [e. g., gravity waves, Rossby waves, Kelvin waves, and Rossby-gravity waves; see Andrews et al. (1987), for further description of these and other types of waves] are often distinguished by the restoring mechanism that is responsible for producing the wavelike motions. In particular, both gravity waves and Rossby waves play a most significant role with respect to the large-scale motions and the concomitant transport of constituents observed in the stratosphere.

The restoring force for the gravity wave is the buoyancy force, which is proportional to N^2 and results from stable density stratification in the atmosphere. One of the most important consequences of gravity waves propagating upward into the stratosphere is their role in determining the structure of the stratospheric jets. Gravity waves propagate upward and "break" (Lindzen, 1981) at some level in the stratosphere or in the mesosphere. Breaking occurs as the gravity wave grows in amplitude, eventually producing an unstable lapse rate with resultant turbulence and mixing. Momentum deposition occurs as a result of the breaking process and effectively creates a net drag on the zonal momentum budget and decelerates the zonal-mean flow. Gravity wave breaking in the lower stratosphere is believed to contribute to the separation of the tropospheric and stratospheric jets as seen in Figure 2*b*. In a similar fashion, gravity wave breaking in the mesosphere contributes to deceleration and closing off above the stratospheric jets in the mesosphere.

The Rossby wave restoring force ultimately results from the latitudinal gradient in the Coriolis parameter, f,

$$f = 2\Omega \sin \phi \tag{9}$$

where Ω is the angular velocity of rotation of Earth and ϕ is the latitude. The Rossby wave is responsible for much of the irreversible quasi-horizontal transport that occurs in the extratropical wintertime stratosphere by a process termed Rossby wave breaking (McIntyre and Palmer, 1984), which is fundamentally different from the gravity wave breaking process just discussed. Contours of constant values of Ertel's potential vorticity, Π, on an isentropic surface (a constant Θ surface) represent material lines when Eq. (5) is valid. Rossby wave breaking occurs as the wave amplifies, and material lines buckle and deform irreversibly. A graphic illustration of the process can be seen in Figures 4*a* and 4*b*. This sequence of figures depicts conditions in the midstratosphere (about 30 km) during a stratospheric warming in January 1979. The shaded area in Figure 4*a* correlates very well with the polar vortex, which was nearly centered on the pole on January 17, 1979. Ten days later (Fig. 4*b*), a large-amplitude wave resulted in elongation of the polar vortex with a tongue of low potential vorticity air from the vortex being drawn around the Aleutian anticyclone (which is centered near 65°N, 150°W at this time), and discrete parts of this tongue are seen to be scattered around the anticyclone. McIntyre and Palmer (1983) constructed these isentropic distributions of Π from meteorological analysis data and refer to them as a "coarse-grained" view of the potential vorticity distribution in the stratosphere.

Concurrent observations of ozone (Leovy et al., 1985) show a very high degree of correlation with the potential vorticity distribution and testify to the essential correctness of the Rossby wave-breaking paradigm of McIntyre and Palmer (1983) for transport and mixing in the extra-tropical stratosphere.

6 SUMMARY

Earth's atmosphere is a thin layer of gas rotating with the planet and externally forced by differential radiative heating by the Sun. The somewhat simplified picture of the large-scale circulation in the stratosphere that has been presented here has been separated into two component parts for convenience, a zonally averaged representation and a contribution from wavelike disturbances. The zonally averaged circulation described here is one that is thermally driven with air parcels heated and ascending at low latitudes, drifting slowly poleward, and finally cooling and descending at higher latitudes. Concurrently, wavelike disturbances propagating upward produce departures from the zonally averaged state. These disturbances transport and mix heat, momentum, and constituents. The reader should be reminded that chemical transformations are also taking place that not only alter the composition of the stratosphere, but also affect the radiative heating as the composition changes (principally changes in ozone, water vapor, and carbon dioxide). It should also be noted that the process by which air enters and leaves the stratosphere (the stratospheric–tropospheric exchange process) is considerably more complicated than described here. A recent comprehensive review of stratospheric–tropospheric exchange can be found in Holton et al. (1995). The stratosphere should be viewed as a very complex system in which radiative, chemical, and dynamical processes mutually interact to determine the structure and composition.

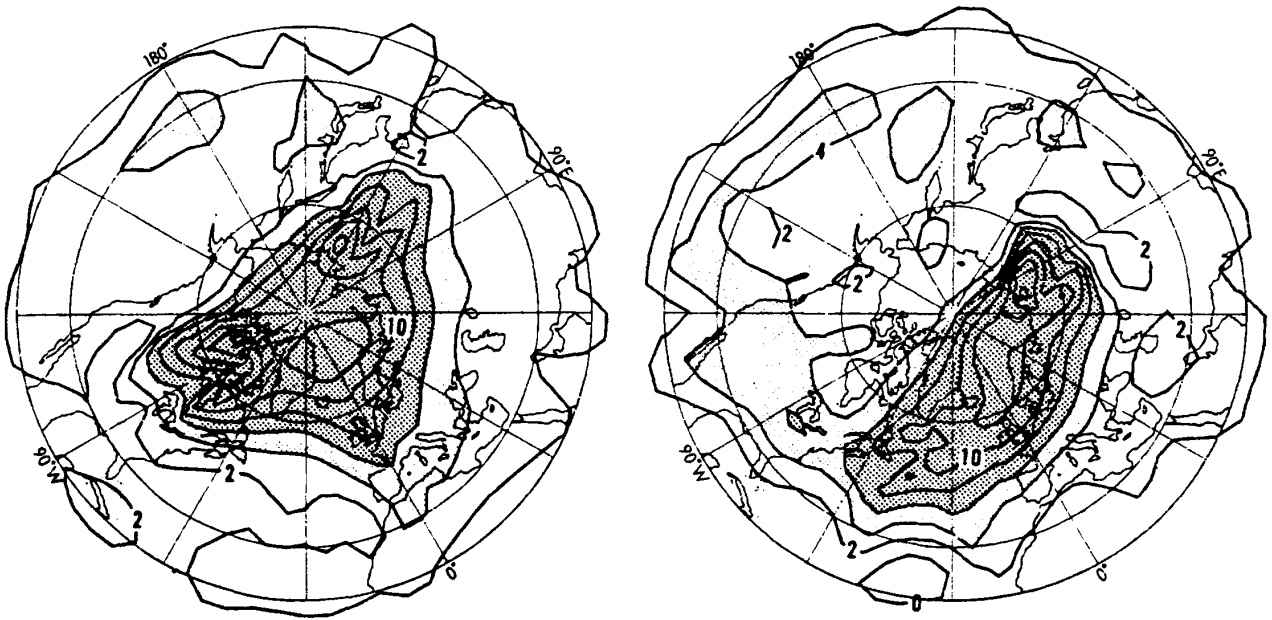


Figure 4 Potential vorticity distribution on the 850 K isentropic surface for (*a*) January 17, 1979 and (*b*) January 27, 1979. Polar stereographic projection with outermost latitude circle 20°N (from McIntyre and Palmer, 1983).

REFERENCES

Andrews, D. G., J. R. Holton, and C. B. Leovy, *Middle Atmosphere Dynamics*, Academic, London, 1987.

Dunkerton, T. J., On the mean meridional mass motions of the stratosphere and mesosphere, *J. Atmos. Sci.*, *35*, 600–614, 1978.

Goody, R., *Principles of Atmospheric Physics and Chemistry.* Oxford University Press, New York, 1995.

Holton, J. R., *An Introduction to Dynamic Meteorology*, Academic, San Diego, CA, 1992.

Holton, J. R., P. H. Haynes, M. E. McIntyre, A. R. Douglass, R. B. Rood, and L. Pfister, Stratosphere-troposphere exchange, *Rev. Geophys.*, *33*, 403–439, 1995.

Lindzen, R. S. Turbulence and stress owing to gravity wave and tidal breakdown, *J. Geophys. Res.*, Vol. *86*, 9707–9714, 1981.

McIntyre, M. E., and T. N. Palmer, Breaking planetary waves in the stratosphere, *Nature*, *305*, (5935), 593–600, 1983.

Murgatroyd, R., The structure and dynamics of the stratosphere, in G. A. Corby (Ed.), *The Global Circulation of the Atmosphere*, Royal Meteorological Society, London, 1969.

Murgatroyd, R., and F. Singleton, Possible meridional circulations in the stratosphere and mesosphere, *Q. J. R. Meteorol. Soc.*, *87*, 125–135, 1961.

Pedlosky, J., *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1979.

Ramanathan, V., and W. L. Grose, A numerical simulation of seasonal stratospheric climate: Part I. Zonal temperatures and winds, *J. Atmos. Sci.*, *35*, 600–614, 1978.

U. S. Standard Atmosphere. U.S. Government Printing Office, Washington, DC, 1976.

Wallace, J. M., and P. V. Hobbs, *Atmospheric Science: An Introductory Survey*, Academic, San Diego, 1977.

WMO Scientific Assessment of Ozone Depletion, WMO Global Ozone Research and Monitoring Project, Report No. 37, World Meteorological Organization, Geneva, Switzerland, 1994.

CHAPTER 21

STRATOSPHERIC OZONE OBSERVATIONS

JACK A. KAYE AND JACK FISHMAN

1 INTRODUCTION

The accurate knowledge of the distributions of ozone (O_3) in the global atmosphere is important for several reasons. First, the amount of ozone in the atmosphere plays a significant role in determining the amount of biologically damaging ultraviolet (UV) radiation that can reach Earth's surface. Second, ozone both absorbs and emits radiation in the atmosphere; this must be accounted for in atmospheric circulation models if they are to correctly represent the temperature and wind distributions in the atmosphere, especially in the upper troposphere and lower stratosphere. Finally, ozone together with the hydroxyl (OH) radical formed in the atmosphere in ozone photochemistry are key atmospheric oxidants. Hydroxyl plays a particularly important role by initiating much of the chemistry associated with air pollution and by being the primary destruction mechanism for several long-lived chemical compounds that contribute to global warming.

Unlike any other atmospheric phenomenon, the U.S. Congress has mandated that the National Aeronautics and Space Administration (NASA) prepare reports describing the status of our current understanding of the upper atmosphere (Public Law 101-549). In accordance with this mandate, several documents have been issued since 1985 in the form of World Meteorological Organization reports; these summaries contain a plethora of information about stratospheric ozone data as well as supporting measurements of other trace gases critical to the destruction of the stratospheric ozone layer. Much of the information in this section can be found in the last few of these reports (Albritton and Watson, 1991; Albritton et al., 1994, 1998), and the reader is referred to these studies for additional in-depth information.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

For the purposes of understanding surface UV radiation, the main quantity for which knowledge is needed is the total column amount of ozone (the integrated ozone amount in a single "column" of the atmosphere). For atmospheric circulation models and both air pollution and chemical oxidation studies, the distribution of ozone as a function of altitude (the vertical profile) must be known. Accurate, long-term knowledge of ozone distributions is important if changes in surface UV flux, upper atmosphere temperature distributions, and concentrations of both long- and short-lived pollutants are to be quantitatively understood.

For ozone distributions to be used in quantitative studies, they must be measured with high accuracy and precision. A particularly important form of precision is long-term measurement stability, as there is strong evidence for long-term changes in ozone distribution (both total column and vertical profile) over much of Earth's surface. Without excellent (and well-characterized) long-term measurement stability, it is difficult to differentiate gradual long-term changes in actual ozone distributions from inaccuracies or drift in the measurement systems.

The measurement of ozone distributions in the atmosphere presents both a challenge and an opportunity. The challenge comes mainly from its limited amount—the amount of ozone in a given volume of air is always quite small, nearly never exceeding a mixing ratio (mole fraction) of 10 parts per million by volume (ppmv) in the stratosphere and on the order of 100 parts per billion by volume (ppbv) in the polluted troposphere. Amounts in the unpolluted troposphere can be significantly smaller than the latter figure (see Chapter 3.) The mixing ratio of ozone can vary significantly with altitude, including a significant increase with altitude on going from the troposphere into the stratosphere, as well as the existence of thin layers (laminae) of air with concentrations of ozone that differ from those of the surrounding air. In regions of severe ozone depletion (such as the lower stratosphere over the Antarctic in the Austral spring), ozone may be nearly completely absent.

The total column amount of ozone typically varies in the range of 3×10^{18} molecules/cm^2 to $\sim 1.5 \times 10^{19}$ molecules/cm^2. This column amount is expressed in Dobson units (DU), which correspond to the thickness of the layer of ozone (in thousandths of a centimeter) that would be formed if all the ozone in a column of air were brought to the surface. The conversion factor is $1\ \text{DU} = 2.69 \times 10^{16}$ molecules/cm^2. The amounts given above correspond to the Antarctic in Austral springtime ($\sim$100 DU) and the Arctic region in late winter ($\sim$500 DU).

Other challenges to ozone measurement can stem from potential interferences. Measurement techniques based on chemical reactivity may be affected by the presence of other oxidizing species, such as sulfur dioxide (SO_2), while spectroscopic measurements may be affected by the presence of species with absorption features in the same wavelength regions as those of ozone. Spectroscopic measurements may also be significantly affected by the presence of aerosol particles, and most such techniques (microwave and far infrared measurements being the exception) cannot penetrate clouds. Finally, it is worth noting that ozone is present in several isotopic forms. The dominant one ($\sim$99%) consists of three atoms of the dominant isotopic form of oxygen atoms (^{16}O), but there are also forms involving ^{18}O and ^{17}O, whose chemical abundance is $\sim$0.3 and $\sim$0.04% of that of ^{16}O.

Although this chapter focuses on the measurement of ozone, it is important to remember that ozone measurements cannot be understood (and especially, the causes of observed changes be understood) if measurements of related parameters, such as temperature, aerosol distributions, and other chemical constituents, are not also made. Indeed, many measurement networks, aircraft-based research platforms, and satellites make several of the measurements together to allow for maximum utility of the information obtained.

2 PROPERTIES OF OZONE AFFECTING ITS MEASUREMENT

The ozone molecule contains three oxygen atoms and is shaped in the form of an isosceles triangle, with a bond angle of 117° and the length of the bond is ~0.13 nm. The electronic spectroscopy of ozone is very rich owing to the multiplicity of electronic states that arise from the combination of a triply degenerate oxygen molecule [$O_2(^3\Sigma_g^-)$] with a ninefold degenerate oxygen atom [$O(^3P)$], as well as the presence of relatively low-lying states of both atomic oxygen [$O(^1D)$] and molecular oxygen [$O_2(^1\Delta)$, $O_2(^1\Sigma)$]. The resulting electronic spectra consist of several important band systems covering the range from the ultraviolet to the near infrared. Figure 1 shows the absorption properties of the ozone molecule in the ultraviolet and visible parts of the electromagnetic spectrum. These strong spectral features—the Hartley and Huggins bands in the ultraviolet and the Chappuis band in the visible—demonstrate the potential for use of ozone spectra in its measurement. In particular, the sharp variation in ozone absorption near 320 nm shows that ultraviolet measurements shortward of this wavelength should be very useful for ozone measurements. The much weaker Chappuis band (near 600 nm) may be useful where long path lengths are available or where ozone amounts are sufficiently high that near saturation could occur with shorter wavelengths.

3 TOTAL COLUMN MEASUREMENTS

Ground-Based Methods

The measurement of the total column amount of ozone in the atmosphere goes back more than 70 years with the development of an ultraviolet technique by Dobson. This technique, still used today, has formed the backbone of all global measurement programs for ozone columns. The basic physics of this technique is relatively straightforward. UV radiation from the sun will be absorbed by ozone in the atmosphere, so ground-based measurements of surface UV flux will contain information about the integrated ozone amount in the atmosphere. As noted in the previous section, other processes, such as Rayleigh and Mie scattering involving atmospheric aerosol particles will also affect ozone measurements. The Dobson technique involves the use of pairs of UV wavelengths corresponding to features with different strengths in ozone's UV spectrum. Since the wavelength sensitivity over a relatively

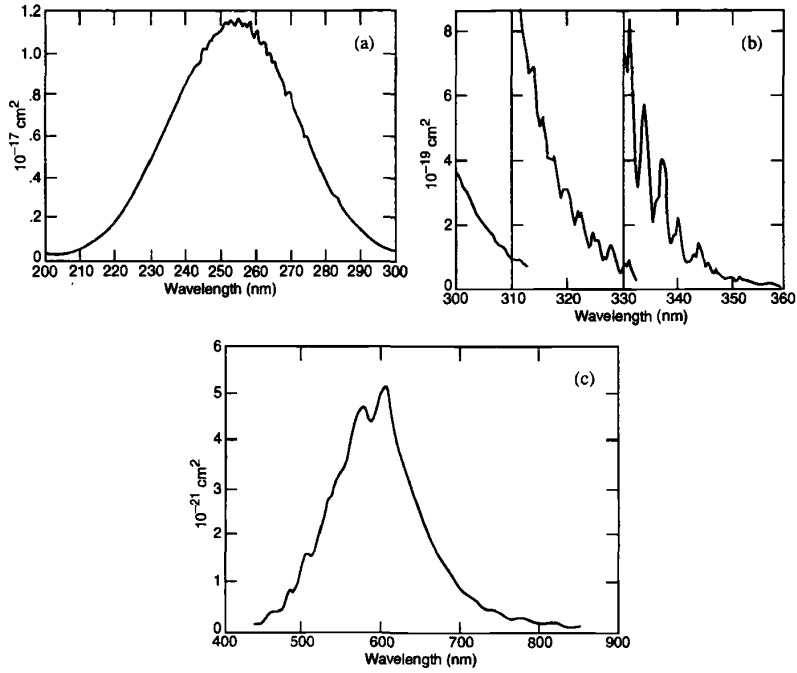


Figure 1 Electronic absorption spectrum of ozone: (*a*) Hartley band and (*b*) Huggins bands in the ultraviolet, and (*c*) the weaker Chappuis band, centered near 600 nm, in the visible.

short spectral region of the scattering by aerosols is much less than that of ozone's UV absorption spectrum, a much improved estimate of total ozone amounts can be retrieved (assuming knowledge of the ultraviolet flux from the sun is available at the two wavelengths used). Different pairs of wavelengths may be used for different amounts of ozone; typically used pairs include 312/331 nm and 318/340 nm.

The distribution of Dobson instruments increased dramatically in the 1950s, and now there is excellent coverage over much of the world with Dobson-type instruments (which included not only those operating on the above principle using a limited number of fixed wavelengths, but also other instruments such as the Brewer spectrophotometer and filter photometers used primarily in the former Soviet Union). Like all surface-based instruments, the Dobson network lacks coverage over much of the ocean-dominated Southern Hemisphere and has fewer stations in developing countries than in industrialized nations. As noted above, such surface-based measurements will not provide data in the presence of clouds, which can be a significant limitation in the tropics, where cloudiness associated with the upward part of the Brewer–Dobson circulation is a common occurrence.

Although the Dobson technique is the most common surface-based one for measurements of the total ozone column, other approaches have been used in the past. These include those using both infrared and visible/UV wavelengths. In the latter, a variant of the Dobson technique, known as differential optical absorption spectroscopy (DOAS) is used. These other techniques have the advantage of not requiring direct sunlight to make measurements, which may be of particular importance in attempts to measure ozone in polar night (when moonlight can be sufficient for ozone measurements).

Space-Based Measurements

The primary space-based measurement technique used for measurements of total column ozone is the backscatter ultraviolet (BUV) technique. This is really a spaceborne analog of the Dobson technique, except when used from space one must account for the fact that the UV radiation passes through the atmosphere at least twice—once on the way from the sun to the surface (or scattering/absorbing layer) and once on its return to the measuring spacecraft. Account must also be taken for the UV reflectivity of the underlying ground or cloud surface. A schematic diagram of how satellite measurements are taken using four different methods is given in Figure 2. The backscatter ultraviolet (as its name implies) utilizes the ozone absorption characteristics in the ultraviolet portion of the spectrum. Occultation techniques use the properties of ozone absorption in the visible and ultraviolet wavelengths whereas the limb emission and limb scattering techniques use a knowledge of ozone absorption in the infrared and microwave portions of the electromagnetic spectrum.

In the BUV technique, the solar flux can be measured directly, although to reduce the flux to manageable levels for the observing instrument, a "diffuser plate" is typically deployed when the instrument looks at the sun (it is retracted for Earth

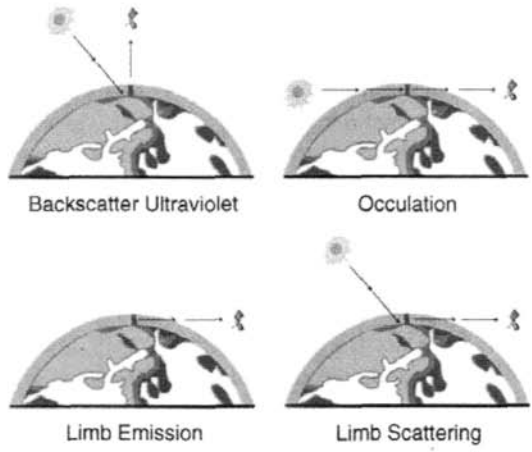


Figure 2 Schematic diagram showing the spacecraft and atmosphere geometry for the four common methods of acquiring trace gas measurements from satellite instruments. See ftp site for color image.

viewing). In applying the BUV technique, it is also helpful to have the measurement time close to local noon, so that the optical path lengths through the atmosphere are close to their potential minimum for the corresponding surface location. In its simplest form, the instrument looks straight down (nadir viewing) to make measurements below the satellite track (Fig. 2*a*). Maps of ozone column can be created by either scanning the instrument's field of view across the orbital track or using some sort of imaging detector so that observations are made corresponding to different ground locations.

The first use of this technique was on the BUV instrument aboard the *Nimbus 4* satellite launched in 1970. This instrument, which was a purely nadir-viewing one, obtained data for several years. The data gave an excellent picture of both the latitudinal and seasonal nature of global column ozone distributions. These data are still of scientific interest; recently they were reexamined to help characterize Antarctic ozone amounts in the early 1970s and show that there was no evidence for significant depletion of Antarctic ozone in the springtime then.

The most significant application of the BUV techniques has been in the total ozone mapping spectrometer (TOMS) and solar backscatter ultraviolet (SBUV) series of instruments. The TOMS instruments use measurements at six wavelengths to measure ozone. For the first two TOMS instruments (one on the *Nimbus 7* satellite that obtained data from October 1978 to May 1993 and one aboard a Russian *Meteor-3* satellite that obtained data from September 1991 to December 1994), the wavelengths used were 312.5, 317.5, 331.2, 339.8, 360, and 380 nm. The latter two are essentially unaffected by the presence of ozone and were used to provide information on surface UV reflectivity. The TOMS instruments were also shown to have information about concentrations of sulfur dioxide, especially during times of enhancement following large volcanic eruptions, aerosols (both tropospheric aerosols including those from biomass burning and volcanic dust, among other types and stratospheric aerosols following large volcanic eruptions), and surface UV radiative flux.

The ground resolution of these instruments is approximately 50 × 50 km at nadir (resolution is degraded as the instrument field of view scans sideways). The orbit of the *Nimbus 7* satellite (sun synchronous, polar orbiting) was excellent for TOMS measurements, while that of the *Meteor-3* satellite was less so since it was not sun synchronous. Roughly half the time the *Meteor-3* orbit led to TOMS observations at local times sufficiently far away from noon that results must be used at great care if at all.

The newer TOMS instruments, which operated aboard the Japanese ADEOS spacecraft (Aug. 1996–May 1997) and NASA's *Earth Probe* (EP) satellite, use a slightly different wavelength set from the previous TOMS instruments. For these new instruments, there is an additional channel to help in the measurement of ozone at high solar zenith angles, as well as a channel to monitor the behavior of the TOMS instrument. The *Earth Probe* TOMS instrument was originally launched into a relatively low (~500 km) orbit to provide for better ground resolution (~26 × 26 km at nadir) than that of the *ADEOS* TOMS instrument (42 × 42 km), which flew aboard the higher orbiting *ADEOS* spacecraft. At the lower altitude, TOMS could not obtain full daily maps over the entire sunlit Earth, however, as

there were interorbit gaps equatorward of approximately 60° latitude. Following the *ADEOS* failure, the *EP* satellite was boosted into a higher orbit (~750 km) to allow for near global spatial coverage.

The SBUV series of instruments includes the original SBUV instrument, which flew aboard the *Nimbus 7* satellite, and updated instruments (SBUV/2) that flew aboard several of the operational meteorological satellites of the American National Oceanic and Atmospheric Administration (NOAA) on the *NOAA-9*, *NOAA-11*, and *NOAA-14* satellites, to date. The SBUV instruments, which also have a capability to determine ozone vertical profile (see Section 4 of this chapter) do not have any cross-track scanning capability and thus do not obtain contiguous daily maps as do the TOMS instruments; they simply obtain data along the daytime subsatellite tracks (the nature of the BUV technique, which requires the presence of sunlight, precludes nighttime data). Long-term calibration information for the SBUV instruments was provided by the Shuttle-borne SBUV (SSBUV) instrument, which flew eight times on the Space Shuttle from 1989 to 1996.

The TOMS and SBUV series of instrument have provided an invaluable database on the total ozone distribution of Earth's atmosphere and its many variations. In Figure 3, a two-dimensional representation (latitude–time) of total ozone distributions derived from the TOMS satellite is shown. Key elements of the total ozone distribution are evident—low total ozone with little seasonal variation in the tropics, highest total ozone values in late winter at high northern latitudes, and lowest total ozone values associated with Antarctic ozone depletion at high southern latitudes during the Austral spring. The long-term changes in the global amount of total ozone determined from *Nimbus-7* TOMS is shown in Figure 4*a*. These data have gone through extensive calibration procedures and comparisons with ground-based Dobson stations to ensure the greatest possible accuracy. The greatest changes have occurred over the Antarctic continent and is seasonal in extent. On the other hand, no statistically significant ozone depletion has been noted in the tropics (see

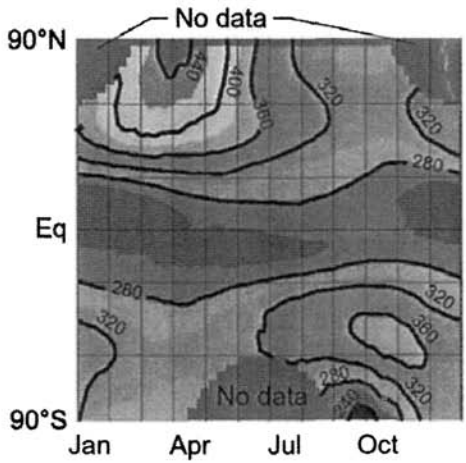


Figure 3 (see color insert) Two-dimensional (latitude/season) representation of total column ozone as measured by TOMS for the period 1978 to 1993. See ftp site for color image.

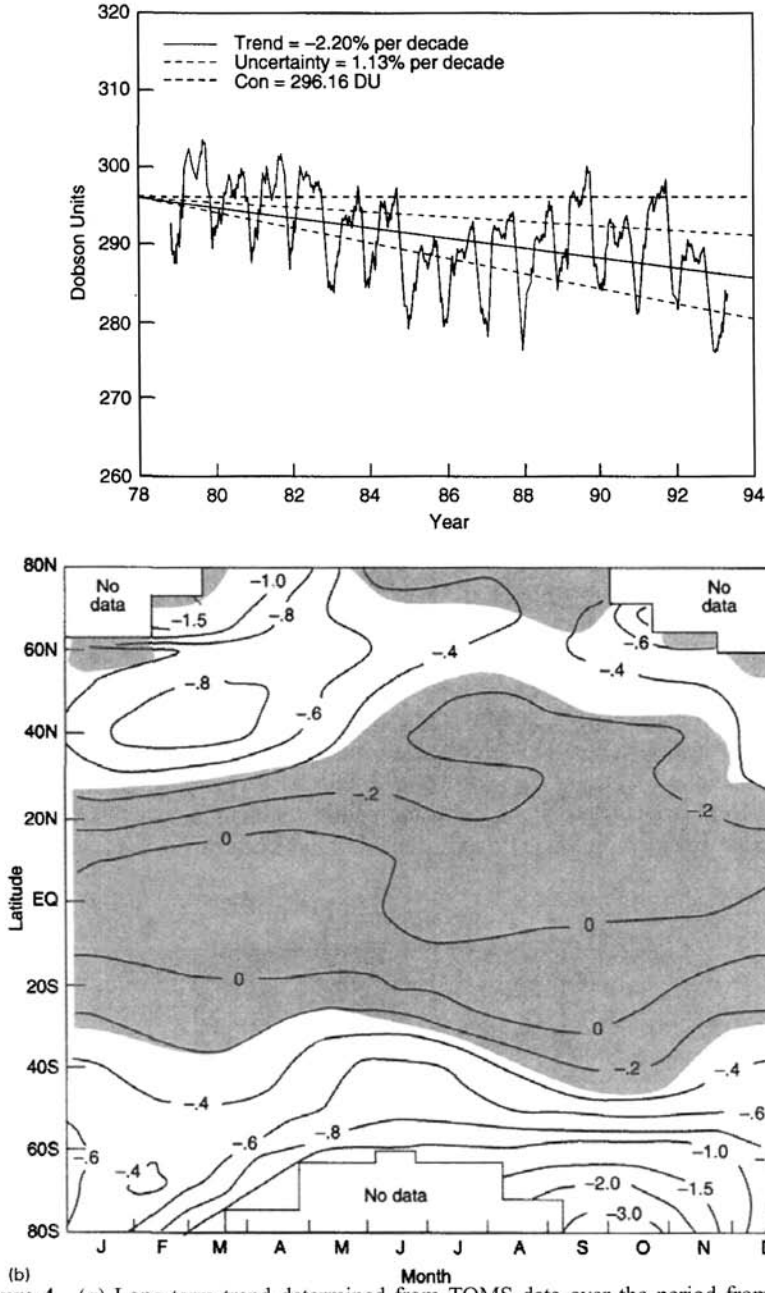


Figure 4 (*a*) Long-term trend determined from TOMS data over the period from 1978 to 1994 after correcting for seasonal cycles, the 11-year solar cycle, and the quasi-biennial oscillation; (*b*) month vs. latitude analysis of ozone trend plotted in contours of percentage loss per year. Shaded areas indicate areas where trend is not definitive. See ftp site for color image.

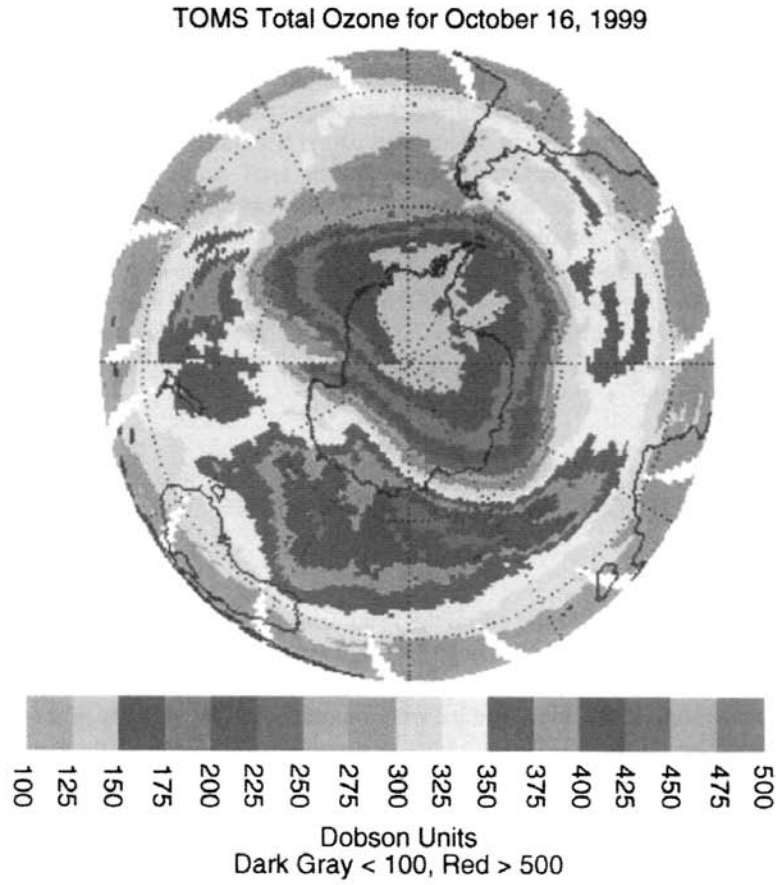


Figure 5 (see color insert) Map of total column ozone over the Antarctic as determined from TOMS October 16, 1999. See ftp site for color image.

Fig. 4*b*). An example of the daily mapping capability of the TOMS instruments is shown in Figure 5, in which a contour map of total ozone distributions during the height of the Antarctic ozone depletion season is presented. The regions of ozone depletion, surrounded by the "collar" region of higher ozone amounts, are clear.

Several limitations of TOMS and SBUV data are worth noting. In particular the UV wavelengths used do not penetrate clouds and are less sensitive to ozone in the lowest few kilometers of the troposphere. Thus, variations in cloudiness or the nature of the tropospheric ozone profile can affect the retrieved ozone column amounts. Improved understanding of these limitations has been an important goal of much recent research.

Another measurement technique used to obtain measurement of a "total ozone-like quantity" is infrared emission. The TIROS operational vertical sounder (TOVS)

instruments measure infrared radiation at 9.6 μm (corresponding to one of the fundamental vibrational modes of ozone). TOVS is mainly sensitive to lower stratospheric ozone and as such does not provide a true total column measurement. However, the correlations between total ozone and lower stratospheric ozone are well established (since it is the lower stratosphere where most ozone is found), so the TOVS product has many of the same characteristics as TOMS total ozone. Since the TOVS measurement uses infrared emission, data can be obtained in regions without sunlight, such as high latitudes during polar night.

A space-based version of the DOAS technique has been implemented aboard the European Space Agency's *ERS-2* satellite using the Global Ozone Monitoring Experiment (GOME). *ERS-2* was launched in April 1995. The GOME instrument uses a broader wavelength range than do the TOMS or SBUV instruments, including longer wavelengths.

4 OZONE VERTICAL PROFILE MEASUREMENTS

Ground-Based Ozone

The first ground-based ozone profiling technique to be used was the Umkehr method, in which the solar zenith angle dependence of Dobson-type measurements is used to determine the vertical profile. This technique obtains data at ~5 km vertical resolution. It provides little information on the lowest ~20 km of the atmosphere, however. In the middle and upper stratosphere, the Umkehr data record has provided an important source of information, especially on long-term ozone trends.

Another ground-based technique for obtaining information on the ozone vertical profile is the use of microwave emission. Since ozone molecules occupy a broad range of rotational states at atmospheric temperatures, there are numerous transitions that will take place for which emission-based remote sensing may be used. Information on the vertical profile comes from the shape of the observed emission lines because of the pressure-broadened nature of the emission lines—emission from ozone in the upper stratosphere will take place near the center of the spectral band, while that from lower down will occur in the wings of the line. Although the vertical resolution of this technique is somewhat limited, it can provide valuable information, especially when measured together with distributions of ozone-destroying free radicals such as ClO or HO_2.

Another technique for ground-based measurement of ozone is lidar. In the lidar technique, a pulse of laser light is sent up from the ground, and the scattered signal that returns to the ground provides information on the composition of the air mass being observed, while the time delay between the laser shot and the return signal is used to provide altitude information. Since the air mass being sampled will interact with laser light by other processes besides ozone absorption (such as molecular Rayleigh scattering, as well as aerosol scattering), lidar systems typically employ two laser wavelengths, one of which is more strongly absorbed by ozone than the other. The wavelength dependence of the aerosol and Rayleigh scattering is typically

much less than that of ozone (and relatively well understood) allowing for retrieval of ozone amounts. The wavelengths used will depend to some extent on the altitude range at which ozone measurements are desired. For stratospheric measurements, where larger ozone abundances are typically observed than in the troposphere, wavelengths with a smaller absorption cross section are needed than for tropospheric measurements. Typical wavelength pairs used are 308 and 355 nm for stratospheric ozone lidar and 288 and 299 nm for tropospheric lidar. Lidar can also be implemented from aircraft, in both upward and downward looking configurations.

In Situ Measurement Techniques

The primary in situ measurement technique used for determination of the ozone vertical profile is that used on ozonesondes. In one standard implementation, an iodine/iodide redox concentration cell is used. An electric current is generated when air containing ozone is pumped into the cell, with the amount of current being related to the partial pressure of ozone in the air mass being sampled. This technique is capable of providing excellent vertical resolution, and is unparalleled at determining the existence of "tongues" or "laminae" of air masses with ozone contents that differ from those of their surroundings (see examples in Chapter 1). Because of the limitations of the balloon on which they are flown, ozonesondes rarely rise above ~30 km. Ozonesonde measurements are usually only made from a fairly limited set of observing stations, and except during certain intensive field campaigns, are typically made at most weekly. Ozonesondes can be flown at various locations and do not require the presence of sunlight. Ozonesondes from the South Pole have provided an important part of our knowledge of the vertical distribution of ozone over Antarctica, for instance, especially on its seasonal variation in springtime. In Figure 6, a plot of ozone vertical profiles over Antarctic measured before and during the presence of the ozone hole are shown. The ozonesondes provide clear evidence for the near total absence of ozone in the 12 to 22 km altitude range during the period of ozone depletion.

The measurement technique used by ozonesondes requires very careful emphasis on calibration and intercomparison. In some cases due to uncertainties about operations, ozonesonde profiles are "normalized to Dobson" so that the observed profile is modified based on a scaling of the calculated integrated ozone column to that observed at a co-located or nearby Dobson station. There are also several different types of ozonesondes, whose operational characteristics differ slightly. In spite of these uncertainties, the ozonesonde record has been critical in the assessment of ozone trends in the lower stratosphere, a region (~15 to 20 km) that is very difficult to observe at high accuracy using space-based instruments.

Other in situ techniques for ozone measurement also exist. One used extensively aboard research aircraft is a spectroscopic technique in which the absorption of air at UV wavelengths is accurately determined. This is a very accurate technique, as the spectral information is well known and there is little opportunity for interference because of the significantly smaller abundance of most potential contaminants. This technique has been used aboard NASA's *ER-2* aircraft in its flights in the lower

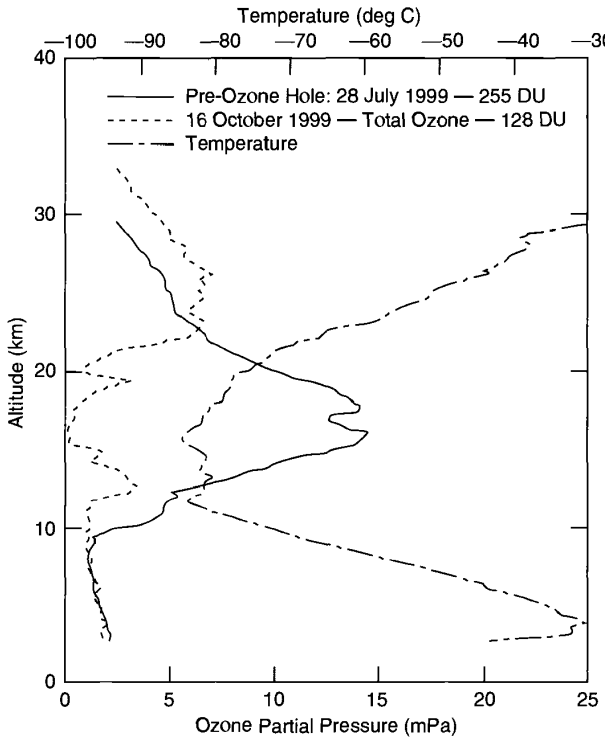


Figure 6 (see color insert) Plot of vertical profile of ozone (blue and red lines) over the South Pole as measured from ozonesondes during austral winter (July 28) and spring (October 16), 1999; temperature profile for October 16 is also shown (green line). See ftp site for color image.

stratosphere and upper troposphere, for instance. A chemiluminescent system has also been used.

Space-Based Remote Sensing

The ozone profiling technique with the greatest heritage uses the BUV technique. By using several wavelengths shorter than those used for total column measurements, information on the ozone distribution in the middle and upper stratosphere can be obtained. This technique makes use of the fact that with decreasing wavelengths light is absorbed at higher altitudes in the atmosphere because of the corresponding increased absorption cross section (see Fig. 1). The vertical resolution of this technique is quite broad, however, some 7 km in the middle and upper stratosphere and close to 10 km below the peak in the ozone layer (typically 20 to 25 km depending on latitude). Through its use on the SBUV series of instruments, this technique has

provided extensive information on the latitude and seasonal dependence of ozone's vertical structure in the middle and upper stratosphere. One strong conclusion to come from this is the clear demonstration of statistically significant ozone losses near 40 km, especially at high latitudes.

The other space-based technique with the longest heritage uses the absorption of radiation at occultation (the rising and setting of the sun with each orbit). The occultation technique (see Fig. 2*b*) has several notable advantages. First, because it involves an along-path length against a rising or setting sun, signals are strong. Second, the technique is inherently "self-calibrating" in that for each measurement an observation is made at the top of the atmosphere and in darkness, so any changes in the performance of the instrument can be determined, at least to first order. Third, the technique has the capability for relatively high (~1 km) vertical resolution to be implemented fairly easily, although in the lower stratosphere, where ozone mixing ratios vary rapidly with altitude and one must view through the peak in the ozone layer, vertical resolution may be degraded.

The main limitation of the solar occultation technique is in spatial coverage. Since the sun rises and sets only once per orbit, at most two latitudes of data per orbit are obtained. These latitudes will change fairly rapidly with time. Depending on the orbital inclination and the orientation of the spacecraft orbit, a complex pattern of observations versus time is obtained; this may complicate the determination of seasonal distributions. If the spacecraft is in a polar sun-synchronous orbit, sunrises and sunsets are only obtained at high latitudes. Another disadvantage of this technique is high sensitivity to aerosol loading. When stratospheric aerosol loading is high, such as following a major volcanic eruption (e.g., Mt. Pinatubo in 1991), the high aerosol abundance provides a great deal of extinction that can complicate the retrieval of ozone amounts. This technique cannot penetrate clouds, and therefore can provide little information on the tropics below the tropopause, as high-level clouds are typically present near the tropical tropopause. An additional limitation of this method is that it actually observes number density as a function of altitude; most atmospheric scientists work with mixing ratios as a function of pressure. Unless temperature is measured together with ozone amounts, the conversion from the observed to desired quantities requires externally supplied (and noncollocated) meteorological information.

The occultation technique has been implemented using both UV/visible/near-infrared and purely infrared wavelengths. The Stratospheric Aerosol and Gas Experiment (SAGE) series of instruments has been the longest term implementation of this technique. SAGE I, which flew aboard the *AEM-2* satellite, obtained data from 1979 to 1981, while the SAGE II instrument, which flies aboard the *Earth Radiation Budget Satellite* (ERBS), has obtained data since its launch in late 1984. Both instruments flew in a 57° inclination orbit; the latitude and time dependence of the solar occultations for SAGE II is shown in Figure 7. It is seen that it typically takes ~3 weeks for the occultations to scan the full range of latitudes. The nonuniform nature of the coverage is quite obvious—in some months, some latitudes are not sampled at all. High latitudes are only sampled occasionally. The SAGE instruments, which make measurements at a total of 4 (SAGE I) and 7 (SAGE II) chan-

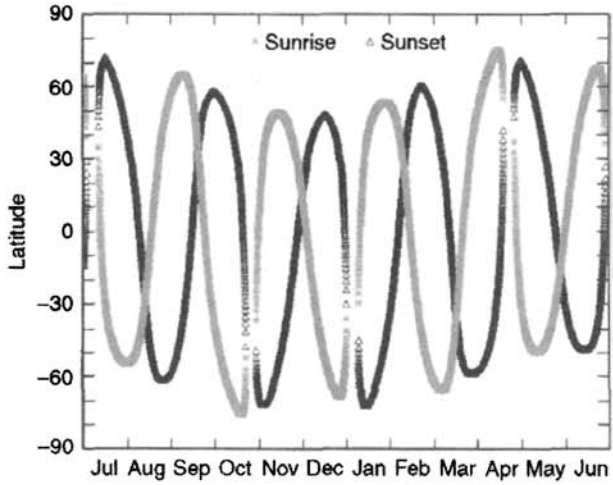


Figure 7 Spatial coverage of solar occultations from the SAGE II instrument as a function of season. Sunrise and sunset occultations are indicated with different symbols. There is excellent year-to-year repeatability of the times and locations of the occultations. See ftp site for color image.

nels, also measured both nitrogen dioxide (NO_2) and stratospheric aerosols; the SAGE II instrument also measures water vapor (H_2O). The main ozone measurement channel uses the Chappuis bands at 600 nm, but the measurement must account for the presence of aerosols, which also contribute to the 600-nm extinction.

Along with ozonesondes, the SAGE II instruments provide a critical data set for the long-term variation of stratospheric ozone. In Figure 8, a zonally and seasonally averaged representation of the long-term trend in stratospheric ozone as a function of altitude is shown. Clear evidence for both upper stratospheric loss in ozone amounts (largest at high latitudes) and lower stratospheric amounts is observed. An accurate characterization of this change as a function of altitude remains an important research area.

The occultation technique using very similar wavelengths was also implemented using the polar ozone aerosol monitor (POAM-2) instrument, which flew aboard the French *SPOT-3* satellite and obtained data for 3 years from late 1993 to late 1996 (Bevilacqua et al., 1997). Since the SPOT-3 satellite was in a polar sun-synchronous orbit, all occultations were at middle and high latitudes, and the POAM data provided an important picture of how high-latitude ozone profiles vary over the course of the year. An example of this is in Figure 9 in which the ozone number density at 20 km is shown inside the Antarctic polar vortex from May through December for the 3 years of POAM-2 observation. The slow decline of ozone in the winter, the rapid falloff in September, and the slow recovery in October and November are all clearly evident. Interannual variability in these 3 years is quite small.

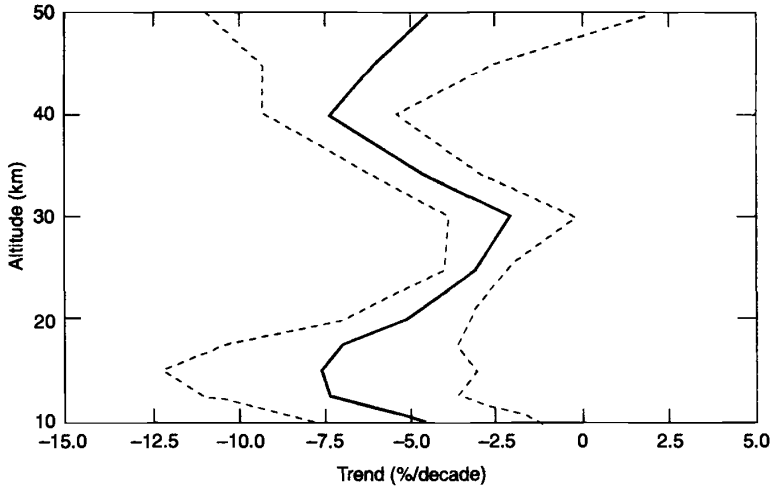


Figure 8 Two-dimensional (latitude/altitude) representations of the long-term trends in stratospheric ozone distributions determined from the SAGE I and SAGE II instruments.

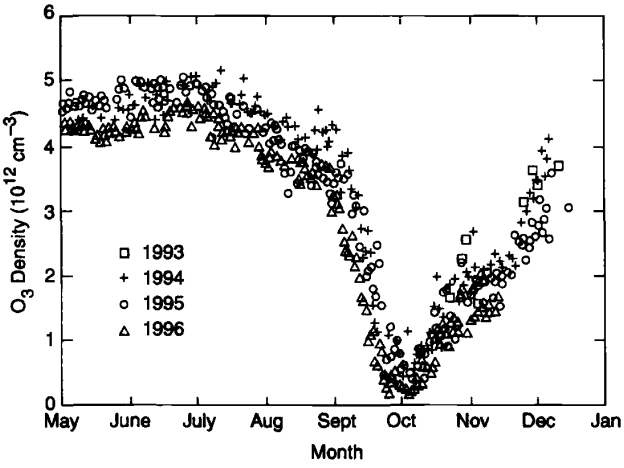


Figure 9 Plot showing the evolution of ozone amounts in the lower stratosphere inside the polar vortex during the fall–winter of the 3 years for which POAM-2 obtained data. Different symbols are used for each year.

The occultation technique has also been applied using infrared wavelengths. The atmospheric trace molecule spectroscopy (ATMOS) instrument, a Fourier transform spectrometer, flew four times aboard the Space Shuttle (1985, 1992, 1993, 1994). ATMOS is notable for its measurements of a very broad range of trace constituents based on its very high spectral resolution ($\sim 0.01\ cm^{-1}$). Its measurement of ozone (and many other constituents) helped serve as an important validation tool for measurements made from NASA's upper atmosphere research satellite (UARS), launched in September 1991. The Halogen Occultation Experiment (HALOE), a combination broadband radiometer and gas cell correlation radiometer has measured ozone (and several other trace gases) during the more than 6 years of UARS operations. Since UARS is in a 57° inclination orbit, the spatial coverage of HALOE is very similar in character to that of SAGE. Most recently, the improved limb atmospheric spectrometer (ILAS) instrument flew aboard Japan's *ADEOS* satellite and obtained data for nearly a year from August 1996 to June, 1997. Since *ADEOS* was in a polar sun-synchronous orbit ILAS data were restricted to high latitudes.

Emission technology has also been used for measurement of atmospheric ozone. These involve looking at the limb of the atmospheres and measuring the thermal emission from ozone (or some other species). Such measurements do not require the presence of a light source and can thus be made over a complete orbit (both day and night). When made from a polar sun-synchronous orbit, they are made at roughly the same time every day (typically once in the daytime and once in the nighttime), which facilitates studies of diurnal variation. On an inclined orbiter, such as UARS, the measurement time will vary and therefore intermix diurnal and seasonal dependence. As typically implemented, limb observations have vertical resolution of ~ 3 km, although the exact amount can vary higher or lower depending on the measuring instrument. Since limb emission is a thermally driven process, simultaneous measurement of temperatures to high accuracy is required. Infrared emission observations of ozone involve the measurement of the emission from the relatively small population of vibrationally excited molecules that exist in thermal equilibrium with the large majority of ground-state molecules, while in microwave emission, it is emission from rotationally excited molecules that is measured. In some cases (especially daytime in the mesosphere), nonthermal process may populate the excited vibrational states of ozone, and these must be accounted for in determining ozone concentrations from infrared emission measurements. Such nonthermal distribution of population states typically does not occur in the microwave, where the smaller energy quantum allows for thermal equilibrium to be maintained.

Implementations of emission techniques for ozone include the limb infrared monitor of the stratosphere (LIMS) instrument, which flew aboard the *Nimbus 7* satellite and obtained data from October 1978 to May 1979. Later implementations include two instruments aboard UARS—the improved stratospheric and mesospheric sounder (ISAMS) and the cryogenic limb array etalon spectrometer (CLAES) instruments. The CLAES instrument, which used a solid cryogen cooler, worked for approximately 20 months until the depletion of the cryogen. The ISAMS instrument on UARS provided 7 months of data.

These instruments have provided significant information on the behavior of ozone at a given pressure level, especially the relationship between ozone amounts and the meteorology of the stratosphere. An example is shown in Figure 10, in which the variation of ozone in the lower stratosphere (~30 mbar) during a major stratospheric warming in the Northern Hemisphere in the winter of 1979 is shown using LIMS data. In this figure, the polar vortex region of high ozone usually found over the pole (e.g., February 6, 1979) is split into two, as shown in the later analyses for February 16 and February 23. The March 1 panel shows that the ozone has filled in the low region subsequent to a stratospheric warming that had occurred during this time.

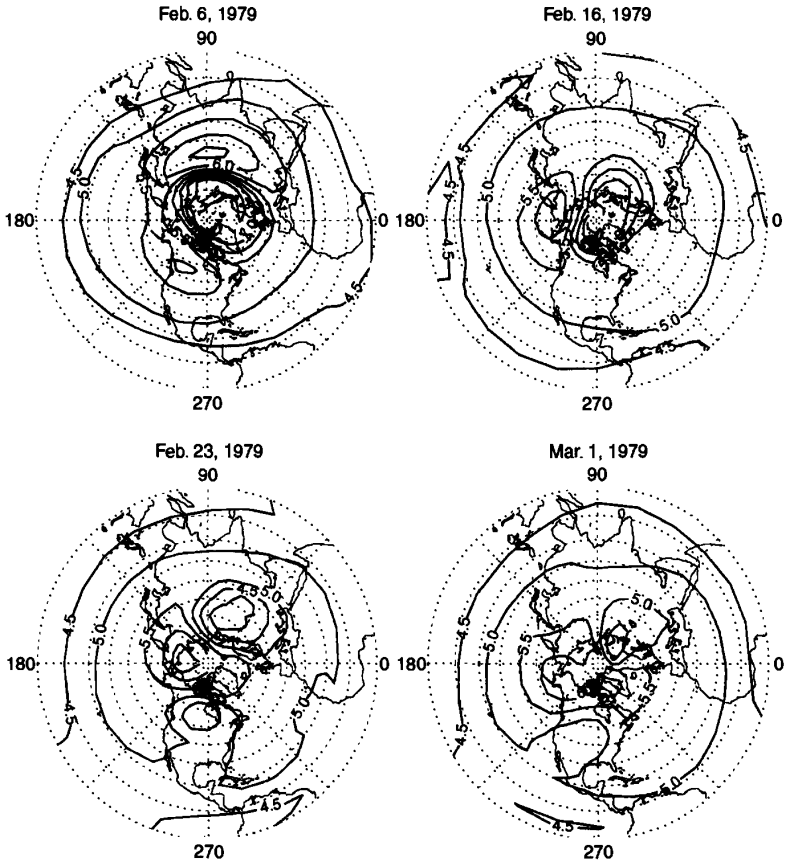


Figure 10 Contour maps showing the evolution of lower stratospheric ozone obtained from LIMS during the vicinity of a major stratospheric warming.

As typically implemented, limb emission observations are made only at longitude as the satellite moves along its orbital track; no cross-track scanning (as is done in TOMS) or imaging is carried out to "fill in" the interorbit gaps. One instrument that is an exception to this is the cryogenic infrared spectrometers and telescopes for the atmosphere (CRISTA) instrument, which was deployed from the Space Shuttle in 1994 and 1997. The CRISTA instrument has three telescopes and infrared detectors viewing 18° apart from each other, thus obtaining higher horizontal resolution than any other atmospheric chemistry profiling instrument.

Several other space-based techniques have been used for measurement of atmospheric ozone. In one, a variant of the infrared emission technique, emission is measured not from vibrationally excited ozone (as was done on LIMS, CLAES, and ISAMS), but from electronically excited molecular oxygen [$O_2(^1\Delta)$] produced following ultraviolet photolysis of ozone. This technique, which was used on the *Solar Mesosphere Explorer* (*SME*) satellite in the early 1980s, is applicable mainly in the thermosphere and mesosphere; at lower attitudes, the $O_2(^1\Delta)$ is quenched so rapidly that there is insufficient signal for detection. UV limb scattering was also implemented on *SME* but was limited to observation in the mesosphere and topmost part of the stratosphere. Since this is a scattering technique, it provides the possibility for good spatial coverage and also can potentially allow for good vertical resolution. More recently, this technique was tested with a pair of instruments (the Shuttle Ozone Limb Sounder Experiment and the Limb Ozone Retrieval Experiment) that flew aboard the Space Shuttle in 1997. Finally, the technique of stellar occultation (in which stars are the source of the radiation) has been tested using the ultraviolet imaging and spectral imagers (UVISI) instrument that flies aboard the U.S. Department of Defense's *Midcourse Space Experiment* (*MSX*) spacecraft. The stellar occultation technique has the potential to overcome the spatial limitation of the solar occultation technique because of the existence of many stars that can serve as a source in a given orbit. The technique also has potential for high vertical resolution; the main complication is the need to overcome the much reduced photon flux for a star as opposed to that of the sun.

5 FUTURE MEASUREMENTS

There will be significant activity in the next few years in ozone measurements, especially in the implementation of several space-based measurement systems. These include additional copies of existing instruments, next-generation instruments based on current techniques and significantly new instruments. The first decade of the twenty-first century will also see to at least two space platforms devoted to studying the composition of the atmosphere and providing new insight into both the distribution of ozone and the chemical processes responsible for its measured abundance.

TOMS/SBUV

An updated version of the TOMS instrument, which will probably have a much wider spectral range than TOMS, as well as a large number of channels, is an instrument to be provided by the Dutch government aboard the *Aura* spacecraft of NASA's Earth Observing System, scheduled for launch in 2004. This instrument, currently known as OMI (ozone monitoring instrument), will also make use of spatial imaging through the use of a detector array, eliminating the need for cross-track scanning and the use of a photomultiplier tube for detection. In the longer-term, a total ozone mapping instrument is scheduled for inclusion in the National Polar Orbiting Environmental Satellite System (NPOESS) being developed by the United States. The exact nature of this instrument has not yet been determined, although it will likely have many of the observing goals of TOMS. The first NPOESS spacecraft will fly no earlier than 2010. Additional SBUV instruments are planned for several NOAA operational meteorological spacecraft through 2004.

SAGE III

An improved version of the SAGE instrument was launched aboard a Russian *Meteor-3M* satellite in late 2001. The SAGE III instrument uses the occultation technique pioneered with the earlier SAGE instruments but has significant advances, including a wider spectral range (from 290 to 1500 nm), and the use of relatively high resolution spectral information within several channels to help provide much improved detection of ozone and other trace species, as well as separation of ozone and aerosol extinction. SAGE III also makes collocated measurements of temperature and pressure to facilitate conversion of profiles from number density versus altitude to mixing ratio versus pressure. SAGE III also has a lunar occultation capability that will remove some of the spatial limitations associated with solar occultations. This is especially important for the *Meteor-3M* SAGE, which will be in a polar sun-synchronous orbit in which solar occultations are restricted to high latitude. The lunar occultations cover a much broader range of latitudes.

Platforms Dedicated to Atmospheric Chemistry

ENVISAT. ENVISAT, a program of the European Space Agency, was launched in 2002 and has three significant instruments for measurements of atmospheric ozone and related trace constituents. These include the scanning imaging absorption spectrometer for atmospheric chemistry (SCIAMACHY), the Michelsen interferometer for passive atmospheric sounding (MIPAS), and the global ozone monitoring through stellar occultations (GOMOS) instruments. SCIAMACHY is an enhanced version of the GOME instrument flying aboard *ERS-2*; it will utilize the DOAS technique as did GOME but will also have limb and occultation modes and will have infrared wavelengths that GOME did not have. MIPAS is a high-resolution infrared

emission instrument, and GOMOS will use the stellar occultation technique to determine ozone profiles (Burrows, 1999).

EOS Aura. The EOS *Aura* spacecraft planned for launch in 2004 will have three ozone-measuring instruments in addition to the OMI. These are the high-resolution dynamics limb sounder (HIRDLS), the microwave limb sounder (MLS), and the troposphere emission spectrometer (TES). HIRDLS will use the technique of infrared emission to determine vertical profiles of ozone, temperature, aerosols, and a host of trace constituents. Unlike most previous infrared emission measurements, however, HIRDLS will have high vertical (~1 km) and horizontal (~5°) resolution; the latter comes from its making several scans away from the nadir as the spacecraft moves in its orbit. MLS is a significant improvement over the UARS MLS, with special attention being given to improving measurements in the upper troposphere and lower stratosphere, as well as the measurement of many trace constituents, such as OH, BrO, and N_2O not measured with the UARS MLS. TES is designed to measure ozone and its precursors in the troposphere using high-resolution infrared spectral measurements. It will use both nadir and limb viewing geometries to do this.

REFERENCES

Albritton, D. L., and R. T. Watson (Eds.), *Scientific Assessment of Ozone Depletion: 1991*, Global Ozone Research and Monitoring Project, Report No. 25, World Meteorological Organization, Geneva, 1991.

Albritton, D. L., R. T. Watson, and J. J. Aucamp (Eds.), *Scientific Assessment of Ozone Depletion: 1994*, Global Ozone Research and Monitoring Project, Report No. 37, World Meteorological Organization, Geneva, 1994.

Albritton, D. L., J. J. Aucamp, G. Megie and R. T. Watson (Eds.), *Scientific Assessment of Ozone Depletion: 1998*, Global Ozone Research and Monitoring Project, Report No. 44, World Meteorological Organization, Geneva, 1998.

Bevilacqua, R. M., et al., Use of POAM II data in the investigation of the Antarctic ozone hole, *J. Geophys. Res.*, *102*, 23643–23657, 1997.

Burrows, J. P., Current and future passive remote sensing techniques used to determine atmospheric constituents, in A. F. Bouwman (Ed.), *Approaches to Scaling of Trace Gas Fluxes in Ecosystems*, Elsevier, Amsterdam, 1999, pp. 317–347

CHAPTER 22

AEROSOL PROCESSES IN THE STRATOSPHERE

MARIO J. MOLINA

1 INTRODUCTION

It is now well established that chemical reactions involving aerosol particles play a key role in stratospheric ozone depletion.[1–3] Some of these reactions take place on the surface of solid particles, while others occur inside liquid particles; both are commonly referred to as heterogeneous processes because they involve both the gas and the condensed phase.

The aerosol layer is located in the lower stratosphere and consists predominantly of aqueous sulfuric acid droplets, commonly labeled SSAs (sulfate stratospheric aerosols). At mid and low latitudes their concentration is 70 to 80% by weight H_2SO_4, corresponding to mole fractions between ~0.3 and 0.5; at high latitudes and in the winter and spring months the SSAs may grow significantly in size, becoming polar stratospheric clouds (PSCs). As they cool, they absorb water vapor and also nitric acid vapor but remain in liquid form, becoming type Ia PSCs. If they freeze, they are labeled type Ib PSCs, and at sufficiently low temperatures they become ice crystals (type II PSCs). The mechanism of conversion between the various stratospheric aerosol types is not well established and is discussed in Section 5 below. Typical particle sizes are ~0.1, 1, and 10 μm diameter for SSAs and PSCs type I and type II, respectively; their abundance is ~1 to 10 particles/cm^3.

2 CHEMICAL REACTIONS ON STRATOSPHERIC AEROSOLS

Stratospheric trace species include sources, free radicals (species with an unpaired electron), and temporary reservoirs. Source species are produced at Earth's surface

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

and are stable enough to eventually reach the stratosphere. Temporary reservoirs are generated in the stratosphere, but they are ultimately transported downwards into the troposphere, where they are rapidly removed by rainout or washout. Both sets of species decompose in the stratosphere producing free radicals—either by photolysis or by reaction with another radical. Free radicals can participate in ozone destruction cycles but can also react with each other to produce stable reservoirs. Thus photochemistry is a source of radicals, while gas-phase reactions interconvert radicals into different forms or else destroy free radicals by producing stable reservoirs. Practically all gas-phase chemical reactions involve free radicals; reactions in the gas phase between nonradical (saturated) species are usually too slow to matter at atmospheric temperatures. However, aerosols provide a pathway for such reactions to take place.

The two most important sets of heterogeneous reactions in the stratosphere are chlorine activation and nitrogen deactivation. Chlorine activation reactions transform temporary chlorine reservoirs to a form that is photolytically active; the most important ones are the following:

$$HCl + ClONO_2 \rightarrow HNO_3 + Cl_2 \quad (1)$$

$$HCl + HOCl \rightarrow H_2O + Cl_2 \quad (2)$$

The species Cl_2 is photolytically very active; it readily absorbs near-ultraviolet (UV) and visible light to produce free Cl atoms, which in turn react rapidly with ozone:

$$Cl_2 + h\nu \rightarrow 2Cl \quad (3)$$

$$Cl + O_3 \rightarrow ClO + O_2 \quad (4)$$

The most important nitrogen deactivation reaction is

$$N_2O_5 + H_2O \rightarrow 2HNO_3 \quad (5)$$

The N_2O_5 species is produced in turn from nitrogen oxides:

$$NO_2 + O_3 \rightarrow NO_3 + O_2 \quad (6)$$

$$NO_2 + NO_3 \rightarrow N_2O_5 \quad (7)$$

The net effect of reactions (5), (6), and (7) is to convert active forms of nitrogen (NO_x) to the relatively stable temporary reservoir HNO_3, a species which, however, may regenerate radicals by solar photolysis:

$$HNO_3 + h\nu \rightarrow OH + NO_2 \quad (8)$$

On the other hand, a significant fraction of the gas-phase HNO_3 is incorporated at low temperatures into PSCs, where it is further stabilized and protected against solar photolysis. Yet another effect of this HNO_3 scavenging process is denitrification: if

some fraction of the PSC particles grow sufficiently large, they may settle to lower altitudes, thus removing the NO_x source more permanently.

The combined effect of chlorine activation and nitrogen deactivation is accelerated ozone depletion: Chlorine free radicals destroy ozone more rapidly in the absence of nitrogen radicals because the two sets of radicals tend to react with each other to produce temporary reservoirs, slowing ozone depletion. The most important example of this process is the formation of $ClONO_2$ (chlorine nitrate) through the following radical recombination reaction:

$$ClO + NO_2 + M \rightarrow ClONO_2 + M \tag{9}$$

Reaction (9) is termolecular, and M is the "third" body (mostly N_2 or O_2) required to stabilize the newly formed bond. In contrast to other radical recombination products such as ClNO (nitrosyl chloride), CONO (chlorine nitrite), and $ClNO_2$ (nitryl chloride), the species $ClONO_2$ is relatively stable toward solar photolysis.[4] However, in the presence of HCl it rapidly regenerates the chlorine radicals by the chlorine activation reaction discussed above [reaction (1)].

3 HETEROGENEOUS REACTION RATES AND MECHANISMS

Rate constants for heterogeneous reaction rates are commonly expressed in terms of a reaction probability γ, which is the probability per collision of a gas-phase reactant molecule with the aerosol surface that chemical reaction will occur. For reaction on a solid aerosol surface of a species with a mean gas-phase concentration [C] molecule/cm^3), the overall rate may be approximated by the following expression, which is based on the "resistance" model[3]:

$$-d[C]/dt = k_t S[C] \tag{10}$$

$$1/k_t = 1/k_{diff} + 1/k_{coll} = R_p/D_g + 4/[\langle v\rangle\gamma/(1 - \gamma/2)] \tag{11}$$

where t is time (s), k_t is the effective overall first-order rate constant (cm/s) for surface reaction, S is the aerosol surface area per unit volume (cm^2/cm^3), $1/k_{diff}$ is the resistance associated with gas-phase diffusion, $1/k_{coll}$ is the resistance associated with molecular collisions with the particle surface, D_g is the gas diffusion coefficient (cm^2/s), R_p is the average particle radius (cm), and $\langle v\rangle$ is the mean molecular speed (cm/s) of the gas-phase reactant. For values of γ of less than ~0.2, the expression $\gamma/(1 - \gamma/2)$ may be approximated by γ. For certain conditions Eq. (11) can be further simplified, e.g., at low pressures and for small particles (large D_g and small R_p), the effect of gas-phase diffusion may be neglected ($1/k_{diff} \approx 0$); etc. On the other hand, for small particles with sizes approaching the gas-phase mean free path, additional correction factors are needed.[5] For reaction on liquid particles, liquid-phase diffusion also needs to be taken into account, leading to additional resistance terms in Eq. (11).[3]

The rate of reaction (5) on sulfuric acid solutions is nearly independent of the acid concentration,[3,4] hence the reaction occurs readily at low and midlatitudes. In contrast, the rates of reactions (1) and (2) are negligible at those latitudes: The reaction mechanism involves as a first step incorporation of HCl vapor into the condensed phase, and HCl is practically insoluble in concentrated H_2SO_4 solutions. On the hand, as the sulfuric acid particles cool and become more dilute at higher latitudes, the solubility of HCl increases sharply, and the reaction probability increases accordingly, reaching values larger than ~0.1 for temperatures below 200 K.

As the particles freeze at high latitudes, the reaction probabilities may be strongly affected: hydrolysis of N_2O_5 [reaction (5)] becomes very slow, while reactions (1) and (2) occur very efficiently on ice surfaces, requiring only a few collisions of the reactant $ClONO_2$ or HOCl with the particles exposed to HCl vapor.[4] Thus nitrogen deactivation [reaction (5)] occurs predominantly at mid and low latitudes and also at high latitudes as long as the aerosol particles remain liquid. In contrast, chlorine activation [reactions (1) and (2)] occurs efficiently on both liquid and solid particles, but only at high latitudes where the temperature drops below a threshold value of about 195 K.

The mechanism of reactions (1) and (2) is ionic in nature: HCl solvates in aqueous phase forming hydrochloric acid, and hence chloride anions. The chlorine atom in HOCl and $ClONO_2$ is slightly electropositive; both of these species react very fast with the chloride anions to produce molecular chlorine, which is rapidly desorbed from the ice surface.

The first step in the mechanism of reactions (1) and (2) on ice particles involves incorporation of HCl vapor into the surface layers. The high affinity of HCl for the ice surface is a consequence of ion pair formation, which takes place because the surface layers of ice are not as ordered as the ice crystal itself; they form a "liquid-like" aqueous layer in the presence of trace amounts of HCl (this species can depress the freezing point of water down to ~195 K). The amount of energy associated with physical adsorption involving only a hydrogen bond is too small to explain the experimental observations of HCl uptake by ice[6]; hence, reaction mechanisms involving weak physical adsorption and resorting to conventional Langmuir-type adsorption isotherms are not suitable.

The second step in the reaction mechanism involves incorporation into the surface layers of HOCl for reaction (2), or $ClONO_2$ for reaction (1). All the reactants have a high mobility on the surface: Once incorporated into the condensed phase, the HOCl molecules almost always find chloride ions before returning to the gas phase. Similarly, the $ClONO_2$ molecules in the surface layers also find chloride ions before reacting with water. This explains the experimental observation of a lack of dependence of the reaction rate on the concentration of HCl vapor: as long as HCl is in excess, the overall reaction is nearly zero order in HCl and first order in HOCl (or $ClONO_2$), and is only very weakly dependent on temperature.

Reactions (1) and (2) also occur rapidly on nitric acid trihydrate (NAT) surfaces, with a mechanism similar to that on ice surfaces. However, there is an additional parameter that should be taken into account, namely the relative humidity. When

NAT is in equilibrium with ice, its H_2O vapor pressure is the same as that of ice, and the reaction probability γ for reactions (1) and (2) is practically the same as that on ice. As the relative humidity (and hence the H_2O vapor pressure of NAT) decreases, the reaction probability γ initially remains high, but it decreases for relative humidity values below $\sim 50\%$ (with respect to ice), and eventually it reaches values more than two orders of magnitude smaller. This behavior can be explained with a reaction mechanism involving the availability of water at the NAT surface to induce solvation and uptake of HCl vapor: at very low relative humidities, there is excess HNO_3 on the surface and solvation is hindered.

4 THERMODYNAMIC PROPERTIES OF STRATOSPHERIC AEROSOLS

To investigate the nature and chemical identity of stratospheric aerosols, it is useful to consider first the thermodynamic properties of the aerosols, and subsequently the rates of transformation between the various phases for the different chemical systems of interest. The primary thermodynamic properties of interest are the mole fractions of the various chemical components of the particles in the condensed phase and their vapor pressures, the partial pressures or concentrations of these components in the gas phase, and the temperature. The vapor pressures for low-volatility components (e.g., NaCl, and sometimes H_2SO_4) need not be considered explicitly; furthermore, the effect of total pressure on thermodynamic properties is negligible for atmospheric conditions.

The thermodynamic properties that determine the stability and equilibrium composition of the various phases can be represented conveniently by phase diagrams. A comparison of the atmospheric partial pressures with the vapor pressures displayed in the phase diagrams provides a useful guideline to establish the identity of the various condensed phases that can exist under atmospheric conditions. A specific atmospheric condition or state can be represented by a point in a phase diagram, while an atmospheric process or trajectory is represented by a line.

To illustrate the use of phase diagrams, consider the H_2SO_4/H_2O system. Figure 1 shows the equilibrium freezing temperatures for this system as a function of composition, and Figure 2 is a logarithmic plot of the water vapor pressure versus inverse temperature. The dashed lines in Figure 2 represent the H_2O vapor pressure of solutions with constant composition; it follows from the Clausius–Clapeyron equation that these lines are nearly straight, their slopes being equal to the partial molar enthalpy of evaporation of H_2O. The solid lines in Figures 1 and 2 represent conditions of coexistence of two condensed phases; the lines separate regions of stability for the different phases. Note also that the stability region for a particular hydrate is represented in Figure 2 by a surface, whereas in Figure 1 it is a vertical line, as the composition of the hydrate is fixed.

Phase diagrams represent properties at thermodynamic equilibrium. However, it often happens that a new phase does not form because of the presence of kinetic barriers to nucleation. This is the case for sulfuric acid solution droplets: They remain liquid throughout most of the stratosphere, even though their temperature

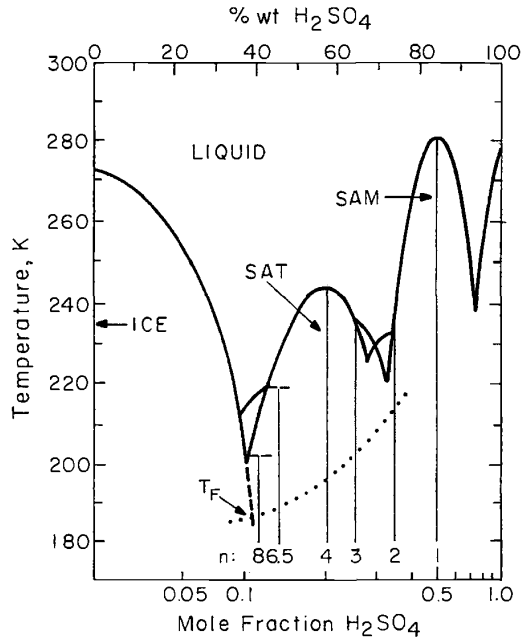


Figure 1 Temperature vs. composition phase diagram for the H_2SO_4/H_2O system. The solid line represents freezing temperatures (solid–liquid coexistence conditions). The thin vertical lines give the composition of the solids. The dotted line represents the equilibrium composition of metastable liquid particles in the stratosphere as a function of temperature, in an air parcel containing 3 ppmv of water vapor at ~16 km altitude; T_F indicates the ice frost point for this air parcel.

is below the freezing point; that is, they supercool very readily. Phase diagrams still provide useful representations of such metastable phases; for example, in Figure 2 the vapor pressures of supercooled solutions are given by extensions of the dashed lines into the solid stability regions. Consider, for example, a stratospheric air parcel around 16 km altitude containing 3 parts per million (ppm) of water vapor, and cooling between 220 and 190 K; the properties of liquid sulfuric acid aerosols in such a parcel are represented by the dotted lines in Figures 1 and 2. The droplets swell and become less concentrated as the temperature drops.

There are similar phase diagrams for the HNO_3/H_2O system; however, because of the relatively high volatility of HNO_3 compared to H_2SO_4, it is useful to consider an additional phase diagram consisting of a logarithmic plot of the HNO_3 vapor pressure versus inverse temperature in order to elucidate the nature of the phases that are stable in the stratosphere for this system. Yet another version of a phase diagram is shown in Figure 3: It is a logarithmic plot with the vapor pressure of one component (HNO_3) in one axis and the vapor pressure of the other component (H_2O) in the

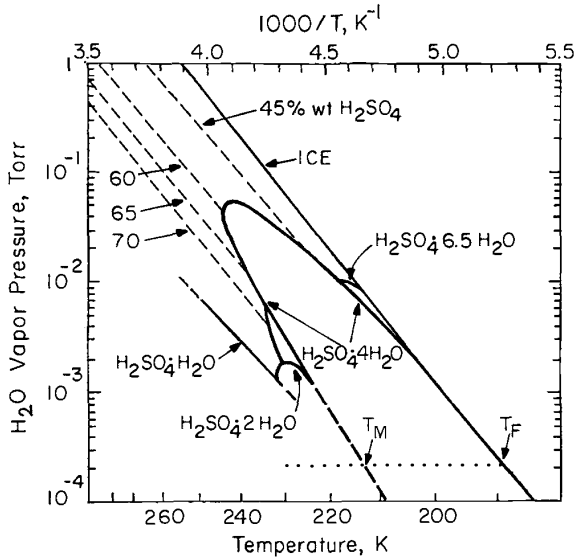


Figure 2 Water vapor pressure vs. temperature phase diagram for the H_2SO_4/H_2O system. The solid lines represent solid–liquid coexistence conditions; the thin dashed lines represent vapor pressures of liquids of constant composition given as wt % H_2SO_4, and the thick dashed line represents equilibrium coexistence conditions for sulfuric acid tetrahydrate (SAT) with the liquid (supercooled with respect to the dihydrate and monohydrate). The dotted line corresponds to that in Figure 1.

other axis. Temperature is a parameter in the figure: Isotherms are straight lines in the solid stability regions, and it follows from the Gibbs–Duhem equation that the slope of these isotherms is related to the composition of the condensed phase: The value of the slope is three for NAT, one for nitric acid monohydrate, and the isotherms are vertical for water ice. The isotherms are curved for the liquid stability region since the composition of the liquid can vary continuously.

Figure 3 shows that the most stable solid phase for the HNO_3/H_2O system in the polar stratosphere is NAT; however, there are indications from laboratory studies that nitric acid dihydrate (NAD) (which is not represented in Fig. 3), may nucleate first,[3] even though for most conditions NAD is metastable with respect to NAT. Note also that Figure 3 shows that the vapor pressure of H_2O over NAT at a particular temperature can have values ranging from the vapor pressure of pure water ice to that of the monohydrate; as discussed above, the reaction probability for reactions (1) and (2) remains large along the isotherm as long as the H_2O vapor pressure does not drop to a value below a half to a third of the value in the ice stability region.[6]

The freezing points and vapor pressure values required to generate phase diagrams such as those shown in Figures 1 to 3 are usually determined directly

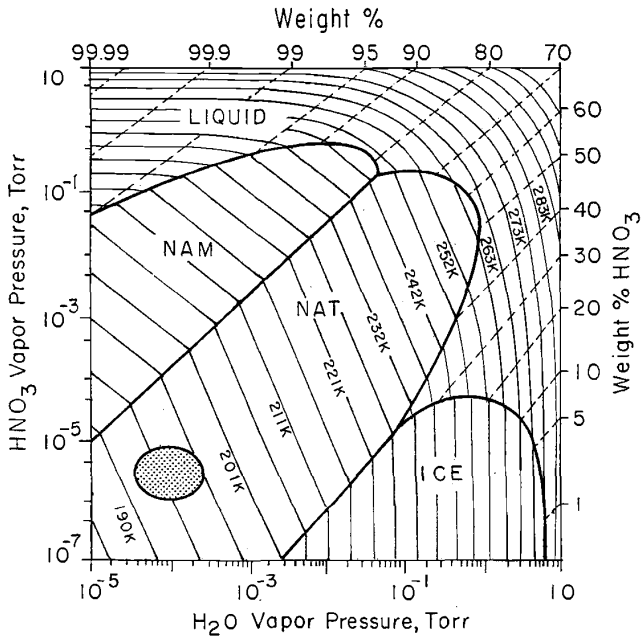


Figure 3 Nitric acid vs. water vapor pressure phase diagram for the HNO_3/H_2O system. The thick solid lines represent coexistence conditions for two condensed phases; the thin lines are isotherms (labeled with T in Kelvin). The dashed lines represent vapor pressures of liquids with constant composition (labeled as wt % HNO_3 in the upper and right axis). The dashed region in the lower left corner represents typical conditions in the lower stratosphere over the poles.

from laboratory experiments. However, the vapor pressures can also be determined indirectly by other means, e.g., from voltage measurements in electrochemical cells—accurate phase diagrams can be constructed from measurements of such voltages together with calorimetric measurements of the enthalpies and temperatures of the various phase transitions of interest. For ternary systems such as $H_2SO_4/HNO_3/H_2O$ the phase diagrams are more complicated but are nevertheless just extensions of the binary diagrams to one more dimension. On the other hand, the vapor pressures for such multicomponent systems can be reliably estimated using semiempirical thermodynamic models.[7]

5 MECHANISM OF FORMATION OF STRATOSPHERIC AEROSOLS

The source of sulfuric acid in the stratosphere is carbonyl sulfide (COS), which is of biological origin. Although emitted from the ground, it is sufficiently stable to reach

the stratosphere, where it oxidizes to form sulfur dioxide, SO_2. This species is further oxidized to produce H_2SO_4 through the following mechanism:

$$SO_2 + OH \rightarrow HOSO_2 \quad (12)$$

$$HOSO_2 + O_2 \rightarrow HO_2 + SO_3 \quad (13)$$

$$SO_3 + 2H_2O \rightarrow H_2SO_4 + H_2O \quad (14)$$

The rate-determining step is reaction (12). Reaction (14) is second order in water vapor[2]; it is fast throughout the atmosphere, except in the upper stratosphere, where the water vapor concentration is relatively small. There is no net consumption of radicals with this mechanism in the atmospheric oxidation of SO_2; the overall effect is merely the conversion of OH into HO_2.

A second important sulfur source consists of volcanic eruptions, a few of which inject SO_2 directly into the stratosphere, such as El Chichón in Mexico, in 1982, and Mount Pinatubo in the Philippines, in 1991. Mount Pinatubo introduced enough SO_2 to increase the stratospheric H_2SO_4 burden by a factor of ~ 30,[8] inducing noticeable global cooling. Satellite observations indicate that the SO_2 oxidation process takes several weeks, and that the excess particles remain in the stratosphere a couple of years; the sulfuric acid haze formed is the origin of bright red sunsets.

As mentioned above, at high latitudes and in the winter months, the sulfuric acid/water droplets cool and grow to become PSCs, absorbing water and nitric acid vapor. Observations in the lower stratosphere of a rapid growth in the volume of these aerosol particles around 195 K were originally interpreted as resulting from the formation of nitric acid trihydrate (NAT); however, a more recent analysis of the field observations indicates that the particles often remain liquid, reaching compositions such as 30 wt % H_2SO_4 and 30 wt % HNO_3[7], with the particle growth being a consequence of rapid H_2O and HNO_3 uptake below a threshold temperature, which happens to approximately coincide with the temperature below which NAT becomes stable.

There is a large nucleation barrier for these supercooled liquid particles to freeze, and laboratory observations show that freezing does not occur until the temperature has dropped several degrees below the ice frost point, which is around 185 K in the lower stratosphere. Under such conditions water ice crystallizes first, leading to the formation of type II PSCs. Some atmospheric observations indicate the presence of solid particles at temperatures above the frost point; it is likely, however, that such particles had reached lower temperatures earlier and that water ice induced the formation of the acid hydrates. As the particles warm up, ice evaporates first and eventually the hydrates melt at the equilibrium phase transition temperatures expected from the phase diagrams (i.e., temperature T_M in Fig. 1), since there is essentially no nucleation barrier for the melting process.

Many questions remain, however, regarding the nature and the rates of liquid–solid phase transformations in PSCs. For example, in the Arctic stratosphere temperatures fall below the frost point much less frequently than over Antarctica, and yet solid PSCs do form, perhaps as a consequence of mesoscale temperature

fluctuations.[9] There are also questions regarding denitrification, the PSC sedimentation process referred to above leads to the removal of nitric acid, and hence, of nitrogen oxides. The process is not sufficiently well understood to permit reliable predictions, for example, of H_2O, HNO_3, and NO_x levels at high latitudes for scenarios involving emissions from proposed future supersonic transports that would fly in the lower stratosphere.

REFERENCES

1. Molina, M. J., L. T. Molina, and D. M. Golden, Environmental chemistry (gas and gas-solid interactions): The role of physical chemistry, *J. Phys. Chem.*, *100*, 12888, 1996.
2. Molina, M. J., L. T. Molina, and C. E. Kolb, Gas phase and heterogeneous chemical kinetics of the troposphere, *Ann. Rev. Phys. Chem.*, *47*, 327, 1996.
3. Kolb, C. E., D. R. Worsnop, M. S. Zahniser, P. Davidovits, C. F. Keyser, M. T. Leu, M. J. Molina, D. R. Hanson, A. R. Ravishankara, L. R. Williams, and M. A. Tolbert, Laboratory studies of atmospheric heterogeneous chemistry, in J. R. Barker (Ed.), *Advanced Series in Physical Chemistry: Progress and Problems in Atmospheric Chemistry*, World Scientific Publishing, Singapore, 1995, pp. 771–875.
4. DeMore, W. B., S. P. Sander, D. M. Golden, R. F. Hampson, M. J. Kurylo, C. J. Howard, A. R. Ravishankara, C. E. Kolb, and M. J. Molina, *Chemical Kinetics and Photochemical Data for Use in Stratospheric Modeling, Evaluation No. 12*, JPL Publication 97-4, Jet Propulsion Laboratory, Pasadena, CA, 1997.
5. Seinfeld, J. H., and S. N. Pandis, *Atmospheric Chemistry and Physics from Air Pollution to Climate Change*, Wiley, New York, 1997.
6. Molina, M. J., The probable role of stratospheric "ice" clouds: Heterogeneous chemistry of the "ozone hole," in J. G. Calvert (Ed.), *The Chemistry of the Atmosphere: Its Impact on Global Change*, Blackwell, Oxford, 1994, pp. 27–38.
7. Peter, T., Microphysics and heterogeneous chemistry of polar stratospheric clouds. *Ann. Rev. Phys. Chem.*, *48*, 785, 1997.
8. McCormick, M. P., L. W. Thomason, and C. R. Trepte, Atmospheric effects of the Mt Pinatubo eruption, *Nature*, *373*, 399, 1995.
9. Carslaw, K. S., M. Wirth, A. Tsias, B. P. Luo, A. Dörnbrack, M. Leutecher, H. Volkert, W. Renger, J. T. Bacmeister, and T. Peter, Particle microphysics and chemistry in remotely observed mountain polar stratospheric clouds, *J. Geophys. Res.*, *103*, 5785–5796, 1998.

SECTION 2

HYDROLOGY

CHAPTER 23

HYDROLOGY OVERVIEW

SOROOSH SOROOSHIAN AND MARTHA P. L. WHITAKER

1 INTRODUCTION

This chapter provides a brief overview of the hydrologic cycle and discusses the role of hydrology, not only in the global contexts of weather and climate but also in the local and regional contexts of weather as it affects water resources management. This chapter contains a description of the hydrologic cycle and the identification of its specific reservoirs and fluxes. In each description, their relevance to various scales of the hydrologic cycle is discussed. The concept of the water balance is subsequently introduced as the basic tool with which one can understand the effects of perturbations on the hydrologic cycle, regardless of the scale of interest. Provided with this knowledge, stakeholders with concerns ranging from global climate change to flood forecasting will be better informed to responsibly manage water resources.

This section of the handbook also contains in-depth discussions on each flux of the hydrologic cycle, including precipitation of rain and snow (Chapters 24 and 25, respectively), evaporation and transpiration (Chapter 26), infiltration and soil moisture (Chapter 27), groundwater flow (Chapter 28), and runoff generation (Chapter 29). The final four chapters describe various types of mathematical tools with which one can analyze hydrologic phenomena: Chapters 30 to 32 specifically describe tools to better understand hydrologic events such as high river flows, runoff, and floods, and Chapter 33 explores the uses of remote sensing and geographic information systems to both visualize and quantify large-scale hydrologic phenomena.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

2 THE HYDROLOGIC CYCLE

Figure 1 represents a conceptual model of the hydrologic cycle and shows Earth's water movement between the ocean, land, and atmosphere. As with all cycles, it is ongoing and continuous, and there is no specific start or end point; however, because the main focus of this handbook is meteorology, precipitation is an appropriate place to begin an evaluation. Precipitation is water released from the atmosphere in the form of rain, snow, sleet or hail. During precipitation, some of the moisture is evaporated back into the atmosphere before ever reaching the ground. Some precipitation is intercepted by plants, a portion infiltrates the ground, and the remainder flows off the land into lakes, rivers, or oceans. An important difference between the roles of snow and rain is that runoff occurs relatively quickly following the rain event, whereas snow usually melts much more slowly over days, weeks, or months. The subsequent surge of snowmelt runoff can provide seasonal recharge to groundwater resources but can also trigger flood conditions if the snowmelt occurs too rapidly and in excessive amounts. In addition, the solid snow or ice may change directly into a gas, skipping the liquid state, in the process called *sublimation*.

When precipitation is intercepted by plants, it is eventually evaporated back to the atmosphere. When it infiltrates the ground, it can be taken up by roots and transpired by plants, it can be evaporated from the soil, or it may recharge an aquifer. The water in an aquifer is called *groundwater*, and its rate of flow in the subsurface is such that water can reside in aquifers for days to centuries before discharging to a surface body of water (e.g., river, lake, ocean). Once groundwater has discharged into a river, lake, or ocean, the surface of the water body is exposed for evaporation, causing moisture to collect and concentrate in the atmosphere, eventually returning to the earth as precipitation as the cycle begins again. In addition to natural discharge, groundwater can more rapidly discharge when an aquifer is pumped. With the advent of motorized pumps, the rapid removal of groundwater from aquifers is a relatively recent phenomenon that has greatly affected the depletion of the aquifers and the water balance of many catchments.

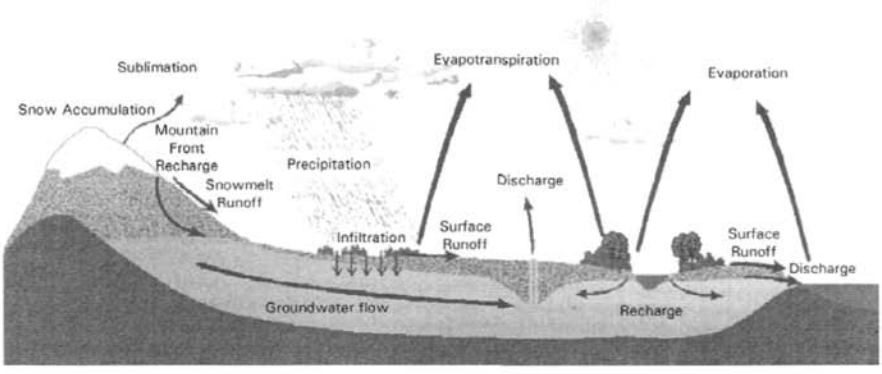


Figure 1 Schematic representation of the hydrologic cycle. (Courtesy B. Imam.)

While the hydrological cycle is a continuous process, it is by no means uniform throughout the globe: the residence time of water varies—often dramatically—among different portions of the cycle. For example, water is continuously evaporated from the surfaces of water bodies (such as oceans, lakes, and rivers). Similarly, precipitation that is intercepted by plants and other surfaces is often evaporated within a matter of hours. Once evaporated, it takes an average of 10 days for a water molecule to cycle through the atmosphere, but if it infiltrates to the water table, or if the precipitation occurs in a polar region, it may reside for hundreds of years before transferring to another step in the hydrologic cycle. In addition to variable residence times, the processes associated with the hydrologic cycle are not evenly distributed over the globe; they vary by climatic region. For example, evapotranspiration occurs readily in semiarid and arid regions, but subsequent precipitation may not occur within the same basin or region. The dramatic differences in how the cycle operates are especially evident when one evaluates the hydrologic cycle at the catchment scale.

Additional, variably detailed discussions of the hydrologic cycle may be found in Horden (1998), Maidment (1993), Driscoll (1986), and Freeze and Cherry (1979). Chahine (1992) offers a particularly thorough discussion of the hydrologic cycle in the context of climate studies and hydrologic modeling.

3 RESERVOIRS

On a global scale, the important reservoirs in the hydrologic cycle are the ocean, atmosphere, polar ice, groundwater, and moisture from land surfaces. At the global scale, water is transferred between reservoirs via four fluxes: precipitation, evapotranspiration, sublimation, and runoff. On a catchment scale, the availability of fresh water is the focus. Critical reservoirs on this scale are the atmosphere, lakes, rivers, and groundwater. Oceans and polar ice are typically irrelevant at the catchment scale, although seasonal snowmelt can contribute significantly (or destructively, in the case of floods) to a basin's water resources. Fluxes within a catchment are more strongly weighted toward the recharge and withdrawal of
potable groundwater, as well as the occurrence of surface water flows.

Fresh water comprises only 2.5% of the world's total water supply. Of this scant freshwater supply, 69.6% is immobilized in ice and snow, primarily in the polar regions; nonsaline groundwater accounts for 30.1%; and the remainder of fresh water (0.3%) is distributed among lakes, rivers, wetlands, atmospheric water, and biological water found in plants and animals. While groundwater is Earth's second largest source of fresh water, on the average it accounts for less than 1% of the earth's total water supply; however, freshwater availability varies greatly on a regional basis. Figure 2 shows the distribution of terrestrial water in terms of Earth's total water supply and Earth's freshwater supply.

The major reservoirs of the hydrologic cycle are described below, and their role in the global- and/or catchment-scale hydrologic cycle is discussed briefly.

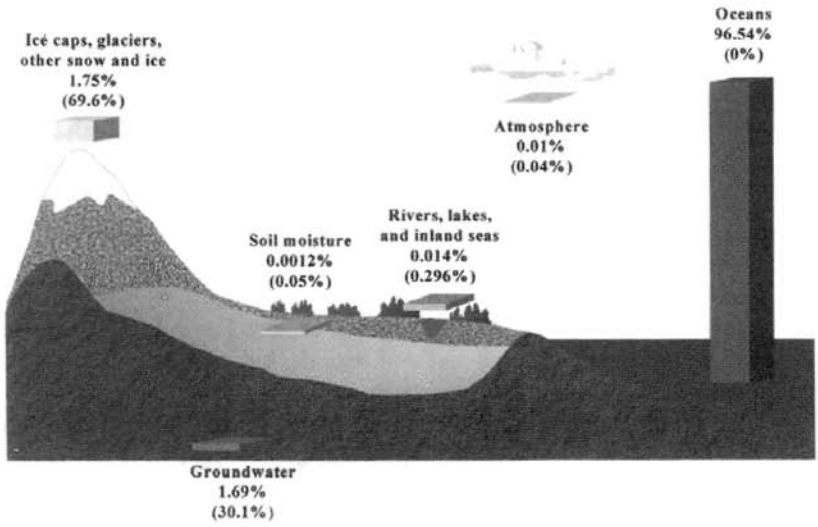


Figure 2 Distribution of terrestrial water partitioned in reference to Earth's total water supply (top percentage) and Earth's freshwater supply (bottom percentage, in parentheses).

Oceans

Roughly 70% of Earth's surface area and 96.5% of Earth's water volume is ocean water. In other words, the oceans comprise 96.5% of the water in the global hydrologic cycle, and with the considerable volume and energy circulated in this vast reservoir, they have a tremendous effect on climate. In particular, ocean surface temperatures (e.g., El Niño Southern Oscillation, or ENSO) can powerfully impact atmospheric circulation patterns.

Polar Ice

While polar ice represents only 1.7% of Earth's total water supply, it comprises almost 67% of Earth's freshwater reserves. Snow and ice-covered surfaces significantly impact Earth's climate because they have a very high ability to reflect solar (short-wave) radiation, but they contribute only marginally to Earth's hydrologic cycle.

Seasonal Snow and Ice

Although seasonal snow and ice represent only 1% of the world's fresh water, the annual melt cycles can play a significant role in water resources management at the catchment scale (Chapter 25). For example, snowmelt can be a welcome source of replenishment for lakes, rivers, reservoirs, and groundwater; however, rapid melting of large volumes of snow can cause flooding and subsequent contamination (e.g., sewer backups) of water resources.

Land-Based Surface Water

Land-based surface water includes rivers, lakes (both fresh and saline), surface soil moisture, and wetlands. On the global scale, the volume of land-based surface water is a small part (0.0153%) of the hydrologic cycle, but the rate of flux through these components, and hence the availability of fresh water, is critically important to human activities within individual watersheds. For thousands of years, humans have interacted with land-based surface water by building aqueducts, digging irrigation canals, and more recently, by diverting or damming rivers, and by pumping, which captures groundwater without allowing it to discharge naturally to a surface water body. With the occurrence of global climate change and the possibility of increased frequency of floods or droughts, responsible hydrologic management will be key to achieving sustainable water resources within catchments. This is possible only through scientists' continuous progress in understanding the various components of the hydrologic cycle in each catchment, improved modeling of all components of the hydrologic system, and narrowing the uncertainty bounds on hydrologic predictions.

Biological water is the primary constituent of living tissue in all plants and animals. It is another form of land-based surface water; yet it is only a minuscule percentage (0.0001%) of the total water on Earth. Regardless of its small percentage, the critical role of plants in the vertical transfer of water from soil and subsurface reservoirs to the atmosphere—particularly in semiarid regions—cannot be ignored (see Chapter 26).

Groundwater

Groundwater generally refers to the water that exists in saturated layers of porous geologic materials, called *aquifers*. On the global scale, groundwater is a slow-moving reservoir that comprises approximately 0.5% of the world's total water, yet it accounts for 30% of Earth's freshwater reserves. Accordingly, the existence and replenishment of these reserves is critical for maintaining a water supply for many human communities.

The global hydrological cycle generally depicts groundwater as slowly discharging to oceans, lakes, and rivers, but groundwater discharge at the catchment scale is rapidly accelerated by the electric turbine pump. A major change in groundwater discharge rates in the United States came about with the widespread agricultural use of electric turbine pumps with the electrification of rural America following World War II. The effect of widespread use of the electric turbine pump for irrigated agriculture has been a decrease in water tables (up to 400 ft within 50 years in fast-growing metropolitan areas such as Tucson and Phoenix, Arizona) and, in some cases, groundwater capture of surface water resources. Only within the past century have scientists and water managers begun to understand the depleting effects of groundwater pumping on surface water flows (e.g., Bouwer and Maddock, 1997; Maddock and Vionnet, 1998; Glennon and Maddock, 1997).

Regional precipitation patterns often determine whether groundwater supplies are a sustainable resource for a catchment's population. In semiarid and arid regions where precipitation is light and water demands increase with growing metropolitan populations, natural recharge is insufficient to maintain a long-term, dependable water supply, and can often result in land subsidence leading to considerable property damage. For some arid regions, the mining of groundwater as a nonrenewable resource is the only viable alternative (e.g., El Geriani et al., 1998; Gijsbers and Loucks, 1999), but in other areas, water conservation and artificial recharge efforts can prolong a basin's water supplies.

Atmosphere

Water resides in the atmosphere for approximately 8 to 10 days before falling back to Earth. The presence of water vapor in the atmosphere affects weather and climate in several ways. First, it is the source for all forms of precipitation (such as rain, snow, sleet, and hail). In addition, it serves to regulate Earth's surface temperature by absorbing and reflecting incoming short-wave solar radiation and absorbing Earth's emission of long-wave radiation. Finally, atmospheric water is a source of latent heat that can mobilize large air masses. Although the average volume of atmospheric water is only 0.001% of Earth's total water reserves, its role in climate and weather is substantial for both global and catchment considerations.

4 FLUXES

The major fluxes within the hydrologic cycle are described below, and their role in the global- and/or catchment-scale hydrologic cycle is addressed.

Precipitation

Precipitation is the process by which liquid and solid-phase aqueous particles, such as rain, snow, sleet, and hail, fall from the atmosphere to Earth's surface. The occurrence of precipitation over land is typically cited as the driving force of the hydrologic cycle, since it triggers the commencement of other fluxes (evapotranspiration, runoff, infiltration) by providing a new source of moisture to the system. The intensity and frequency of precipitation vary considerably both spatially and temporally, and the effects of precipitation can be both welcome (e.g., during droughts) or undesirable if it occurs in excess and causes subsequent flooding. In some regions, where dry air dominates the weather conditions, precipitation may fall from the clouds but evaporate before ever reaching the ground; this is a phenomenon known as *virga*. Measurements and estimates of precipitation (volume and intensity) are critical to any study or modeling effort involving the hydrologic cycle. Rain gages have been the primary mechanisms for observation, but their sparse distributions and other limitations do not provide the spatial and temporal resolution needed for various modeling and research efforts. Recent advances are rapidly improving the

situation by merging satellite and radar with gage information. Examples of such work are discussed in Smith (Chapter 24), Bales and Cline (Chapter 25), Sorooshian et al. (2000), Adler et al. (1993), Arkin and Xie (1994), and Xie and Arkin (1995, 1996, 1997).

Evapotranspiration

Evaporation is defined as "the rate of liquid water transformation to vapor from open water, bare soil or vegetation with soil beneath" (Shuttleworth, 1993), and *transpiration* is the rate of water added to the atmosphere as it moves from soil through the stomata of vegetation. Evapotranspiration (ET) is thus a compound term that describes the collective effect of evaporation of water and transpiration of plants. It is the primary process that moves moisture from Earth's surface to the atmosphere. The only other natural means by which water is transferred from the earth to the atmosphere is the process of *sublimation*, where solid phases of water (e.g., snow and ice) transition directly to atmospheric vapor in the absence of melting. Sublimation typically occurs in regions of cool temperatures and low relative humidity. Evapotranspiration is often an elusive variable to quantify, as it varies diurnally, seasonally, and with changes in precipitation events. A more thorough discussion of evaporation, including a description of various evaporation measurement techniques, may be found in Chapter 26.

Runoff

Runoff is generally thought of as the movement of excess rainfall across the land surface into rivers, lakes, or the ocean. It occurs when the rate of precipitation exceeds the rate of infiltration at the soil surface, or when soil is saturated. Runoff is a particularly important process at the catchment scale, since it can recharge reservoirs and replenish rivers that may subsequently recharge the groundwater; runoff can also cause soil erosion, and excess runoff can lead to flooding. In Chapter 29, Beven offers a broader historical description of the definition of runoff and also describes various hydrological components that contribute to its generation.

Groundwater

Natural groundwater fluxes are typically slow; water may reside in an aquifer for as little as a few hours (as in the case of river bank storage) or for hundreds of years. Accordingly, groundwater itself is often perceived, on the average, as a relatively slow-moving reservoir in the global hydrologic cycle. At the catchment scale, however, where stream–aquifer interactions are relatively rapid and substantial, the average groundwater fluxes are relatively fast moving. They comprise: (1) the natural flow of water between watersheds, (2) the water pumped from an aquifer, (3) mountain-front recharge (seasonal infiltration of snowmelt at the base of mountain ranges), (4) event-based infiltration (infiltration from precipitation and subsequent rises in surface water levels, especially rivers), and (5) artificial recharge via anthro-

pogenic conservation projects. To better comprehend such complex hydrologic flow scenarios, it is critical to first understand the basic principles of groundwater flow. In Chapter 28, Yeh not only describes Darcy's law, the fundamental flow equation for fluid in porous media, but also reviews flow equations for various aquifer conditions (e.g., confined, leaky, unconfined) and describes the use of groundwater flow models used for water resources management.

The Water Balance: Global to Catchment Scale

The *water balance* simply refers to the volumes of water that flow through various components of the hydrologic cycle. More specifically, it is another useful conceptual model in which the components of the hydrologic cycle are evaluated as storage units that are affected by various inputs and outputs. If the various components of the cycle can be quantified or at least estimated, it is possible to gain an understanding of how alteration of a component might affect the balance of the hydrologic cycle. The most simplistic formulation of a water balance is denoted by the elementary continuity equation that conveys the notion that "input to a hydrologic system equals the output from the system, plus or minus any changes in storage":

$$I = O \pm \Delta S$$

where, for a given domain, I is the total inflow, comprised of surface runoff (into the domain), groundwater inflow and precipitation; O is the outflow of evapotranspiration, surface runoff (out of the domain), and groundwater; and ΔS is the change in storage, whose variables are determined by the scale of the domain.

The concept of a water balance is useful at both global and regional scales. At the basin or watershed scale, where groundwater–surface water interactions might encompass the primary focus, precipitation and groundwater inflow would be a model's input, while overland flow, groundwater outflow, and evapotranspiration

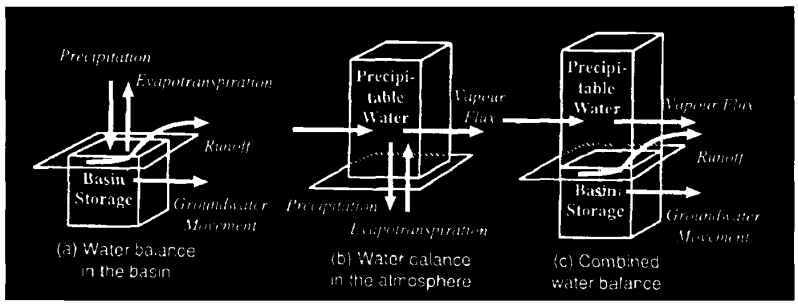


Figure 3 (*a*) Water balance at the catchment scale, (*b*) atmospheric water balance, (*c*) combined land surface atmosphere water balance (after Oki, 1995, 1999).

would be its outputs. Figure 3*a* shows a conceptual model for the water balance at a watershed scale. Note that the figure does not represent changes in storage caused by anthropogenic activities such as pumping, artificial recharge, or surface water diversions from or to other basins. The consideration of such anthropogenic effects in determining the water balance may be critical, depending on the spatial and temporal scales under consideration.

For meteorologists, the most relevant transfers of water in the hydrological cycle are the vapor flux and moisture exchanges between the atmosphere and Earth. With Earth's surface as the focal point of the cycle, we can evaluate precipitation as the major input to the system, while evaporation and transpiration output moisture to the atmosphere. The change in storage could include, for example, the infiltrated water that is not reevaporated into the atmosphere, or water that becomes frozen in polar ice caps. Such water is temporarily and relatively static in the land surface–atmosphere system. That is, given that atmospheric scientists tend to evaluate the hydrologic cycle over a time frame of about 8 to 10 days (i.e., the average amount of time that water cycles through the evaporation–condensation–precipitation cycle), water that resides as ice or becomes slow-moving groundwater is seen as a very slow change in storage of the land surface–atmosphere system. Figures 3*b* and 3*c* show conceptual models of the water balance in the atmosphere and the combined basin–atmosphere water balance, respectively.

In addition to catchment-scale analyses, the evaluation of the water balance of the hydrologic cycle at increasingly larger scales is important because the issues and

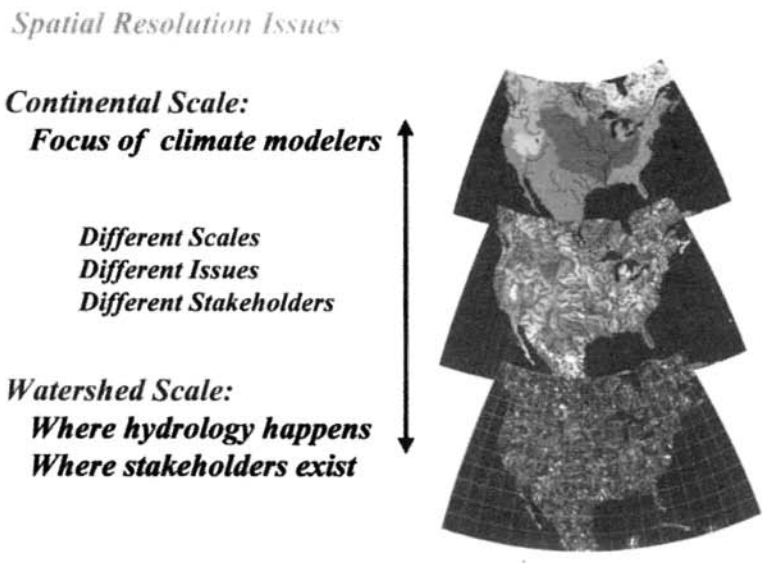


Figure 4 Identification of the spatial scales at which hydrologic phenomena are measured. Different scales delineate different stakeholders, and also determine the various levels of water management issues.

stakeholders are different for each scale. Figure 4 illustrates the different stakeholder affiliations at various scales of hydrologic investigation. The uppermost illustration of North America shows outlines of continental-scale basins, and the adjacent text indicates that the corresponding stakeholders are climate modelers. The results of climate modelers' research potentially affect international, global policies and may influence industrial emissions standards for greenhouse gases. In the middle illustration of North America, sub-basins are delineated, and in the bottom illustration, copious individual watersheds are outlined. The sub-basin to watershed scales are where hydrology happens on scales at which most people can observe more immediate and obvious impacts to their local water supply. Water resources management issues at the watershed scale are thus clearly different than those of the sub-basin and continental basin; however, our understanding of the water cycle at all scales—both spatial *and* temporal—is critically important to addressing the needs of various stakeholders.

Figure 5 shows how specific water resource issues vary in space and time. The spatial scale varies from 10 to 10^6 km^2, and the temporal scale varies from days to centuries. Across the top of the diagram, the types of prediction are identified that can be made for a corresponding time scale: the shortest predictions are weather forecasts (ranging from less than a day to several days), while the longest predictions are climate change (on the order of centuries). The center of the diagram identifies the types of water resources management issues corresponding with different permutations of spatial and temporal ordinates.

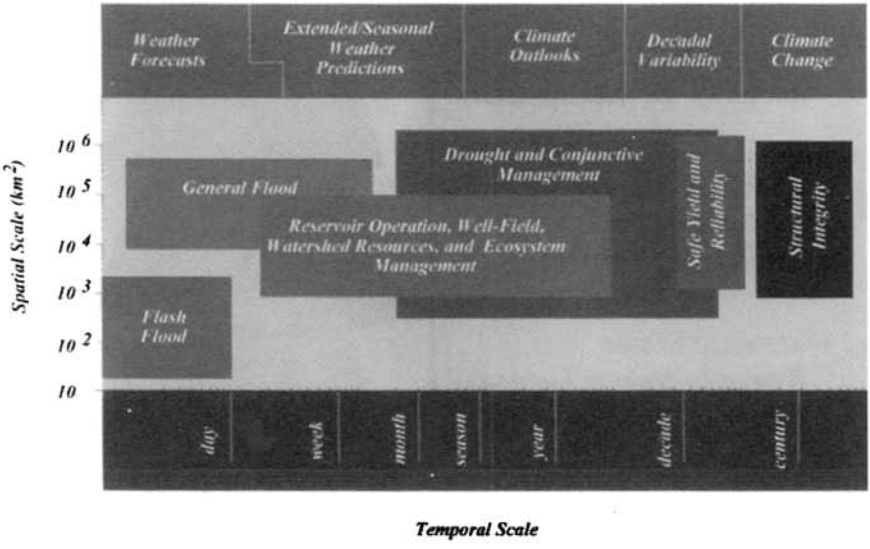


Figure 5 Schematic illustration of how water resources issues vary across spatial and temporal scales (after National Research Council, 1998).

5 MODELING AND REMOTE SENSING OF THE GLOBAL HYDROLOGIC CYCLE: MODELING GLOBALLY, BENEFITING LOCALLY

Modeling efforts can help us understand the potential impact of human activities on the hydrologic cycle and climate in particular. Unfortunately, our current efforts are hampered by a lack of quantitative data on the distribution and flux of water in its various states, and by our uncertainty of the interactive functioning within the hydroclimatological system (Chahine, 1992). To address these uncertainties, the Global Energy and Water Cycle Experiment (GEWEX) was initiated by the World Climate Research Programme (WCRP) in 1988, to observe and model the hydrological cycle and energy fluxes in the atmosphere, at the land surface, and in the upper oceans. A complementary, parallel program, known as the Biospheric Aspects of the Hydrologic Cycle (BAHC), was initialized by the International Geospheric-Biosphere Programme (IGBP) to complement the GEWEX program by placing the emphasis on the biological aspects of the hydrologic cycle—particularly the role of plants in the vertical transfer of water and carbon between the land and atmosphere. For its part, GEWEX has served as a coordinating body of scientists who initiate and facilitate communication among numerous international research teams investigating various aspects of hydrometeorological processes. The hydrological cycle between the land surface and upper atmosphere has subsequently received considerable attention (Chapter 27). Scientists have begun to suggest that we should also consider how land–atmosphere interactions at the basin-scale affect or are affected by climate.

The National Research Council (1998) stated that most water resources management problems are addressed at the sub-basin and watershed scales. Five GEWEX continental-scale experiments (CSEs) have been making promising contributions to improving our understanding of the water balance at scales small enough to be useful for water resources management purposes. For example, the first CSE to be established was GCIP, the GEWEX Continental-Scale International Project, a large-scale study of the Mississippi Basin. During its early phases, GCIP developed data sets, models, and a research framework to better understand and predict land–atmosphere interactions on climatic time scales (seasonal and annual) in the Mississippi River Basin. In fact, GCIP succeeded in meeting most of its objectives, and the project has since transformed to encompass the entire continental United States, as well as part of northern Mexico. This follow-on research project is called the GEWEX America Prediction Project (GAPP). Additional CSEs were selected to represent different climatic conditions than in the Mississippi River Basin. Evaluated together, and separately, the resulting coupling of land surface models with atmosphere and ocean models is a primary step toward improved climate prediction (Chahine, 1992). Such improvements of operational hydrologic and water resources management tools are critical in helping to bring global and GCIP/GEWEX-scale climate predictions down to a scale important for addressing local and regional water resources issues (National Research Council, 1998).

With the ever-increasing popularity of geographic information systems (GIS) and remote sensing (RS), we are witnessing many new advances in hydrologic modeling,

particularly distributed models, which more accurately represent spatial features. Engman and Mittikalli (Chapter 35) provide a brief summary of GIS and RS issues.

6 STOCHASTIC MODELS OF HYDROLOGIC PROCESSES

A *stochastic* process is described by a randomly determined set of observations, each of which is a sample of one element from a probability distribution. Virtually all hydrologic processes can be characterized as stochastic. It is therefore not surprising that the development and application of statistical and stochastic methods in hydrology date back several decades (e.g., Fiering (1967, 1976); Haan (1997); Chow et al. (1998), among many others). The application of flood frequency analysis in hydrologic design and operation of water resources systems is a good example of how influential and powerful these methods have become. Valdés et al. (Chapter 34) address the methods used for stochastic *forecasting*, while Salas et al. (Chapter 33) discuss stochastic *simulations* in the context of precipitation and streamflow. Forecasts are generally applied to operational and management scenarios, while simulations are used in the context of design and planning. More recently, it has become increasingly popular to apply stochastic simulation tools to more thoroughly address the uncertainties of hydroclimatic processes. Salas and Pielke (Chapter 32) provide an excellent review of the current state of the literature in this area.

7 CONCLUSION

The discussion provided above is a brief overview of the various elements of the hydrologic cycle (fluxes and processes) and also offers a summary of ongoing related research activities. It is expected that research and development activities in hydrology and water resources will continue to follow two general paths: theoretical and applied. Theoretical research most related to this handbook will be driven by the need to more accurately close the water budget and quantify the energy cycle at various spatial and temporal scales. As discussed in the chapters that follow, we expect to see future advances in observational tools and applications (e.g., remote sensing, GIS, etc.). We may also expect more advanced modeling of hydrologic processes both at the catchment scale as well as scales that are intended to provide coupling with other components of Earth's systems (i.e., atmosphere, ocean, and biogeochemical processes). On the more applied side, the future requirements for adequate water supplies (quantity and quality) will demand further development of both deterministic and stochastic tools that take advantage of more advanced forms of observational, GIS, and computational techniques. These tools will provide prediction and simulation capabilities for assessing the ramifications of hydroclimatic scenarios (e.g., droughts and floods, regional groundwater depletions, existence and movement of contaminants in both surface water and groundwater, etc.).

Acknowledgments

We gratefully acknowledge the support provided by SAHRA (Sustainability of Semi-Arid Hydrology and Riparian Areas), an NSF Science and Technology Center at the University of Arizona, as well as the Global Energy and Water Cycle Experiment (GEWEX). Our sincere gratitude is also extended to Terri Hogue, Thomas Pagano, and Corrie Thies, for their thoughtful comments and editing of this document in its various stages of completion.

REFERENCES

Adler, R. F., A. J. Negri, P. R. Keehn, and I. M. Hakkarinen, Estimation of monthly rainfall over Japan and surrounding waters from a combination of low-orbit microwave and geosynchronous IR data, *J. Appl. Meteor.*, *32*, 335–356, 1993.

Arkin, P.A., and P. Xie, The global precipitation and climatology project: First algorithm intercomparison project, *Bull. Am. Meteor. Soc.*, *75*, 401–419, 1994.

Bouwer, H., and T. Maddock III, Making sense of the interactions between groundwater and streamflow: Lessons for water masters and adjudicators. *Rivers*, *6*(1), 19–31, 1997.

Chahine, M. T., The hydrological cycle and its influence on climate, *Nature*, *359* (Oct. 1), 373–380, 1992.

Chow, V. T., D. R. Maidment, and L. W. Mays, *Applied Hydrology*, McGraw-Hill Higher Education, New York, 1988.

Driscoll, F. G. (Ed.), *Groundwater and Wells*, Johnson Division, St. Paul, MN, 1986.

El Geriani, A. M., O. Essamin, P. J. A. Gijsbers, and D. P. Loucks, Cost-effectiveness analysis of Libya's water supply system, *J. Water Resourc. Plan. Manag.*, *124*(6), 320–329, 1998.

Fiering, M. B., *Streamflow Synthesis*, Harvard University Press, Cambridge, MA, 1967.

Fiering, M. B., Reservoir planning and operation, in H. W. Shen (Ed.), *Stochastic Approaches to Water Management*, Vol. 2, Ft. Collins, CO, 1976, pp. 17:1–17:21.

Freeze, R. A., and J. A. Cherry, *Groundwater*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

Gijsbers, P. J. A., and D. P. Loucks, Libya's choices: Desalinization or the great man-made river project. *Phys. Chem. Earth (B)*, *24*(4), 385–389, 1999.

Glennon, R. J., and T. Maddock III, The concept of capture: The hydrology and law of stream/aquifer interactions, in *Proceedings of the Forty-Third Annual Rocky Mountain Mineral Law Institute*, Denver, CO, 1997, Chapter 22.

Haan, C. T., *Statistical Methods in Hydrology*, Iowa State University Press, Ames, Iowa, 1977.

Horden, R. H., The hydrologic cycle, in R. W. Herschy and R. W. Fairbridge (Eds.), *Encyclopedia of Hydrology and Water Resources*, Kluwer Academic Publishers, Dordrecht, pp. 400–404, 803 p. The Netherlands, 1998.

Maddock III, T., and L. B. Vionnet, Groundwater capture processes under a seasonal variation in natural recharge and discharge, *Hydrogeol. J.*, *6*, 24–32, 1998.

Maidment, D. R., Hydrology, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, Chapter 1.

National Research Council, *GCIP Global Energy and Water Cycle Experiment (GEWEX) Continental-Scale International Project: A review of progress and opportunities*, National Academy Press, Washington, DC, 1998.

Oki, T., K. Musiake, H. Matsuyama, and K. Masuda, Global atmospheric water balance and runoff from large river basins, *Hydrol. Proc.*, *9*, 655–678, 1995.

Oki, T., The global water cycle, in K. A. Browning and R. J. Gurney (Eds.), *Global Energy and Water Cycles*, Cambridge University Press, Cambridge, 1999.

Shuttleworth, W. J., Evaporation, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, Chapter 4.

Sorooshian, S., K.-L. Hsu, X. Gao, H. Gupta, B. Imam, and D. Braithwaite, Evaluation of PERSIANN system satellite-based estimates of tropical rainfall, *Bull. Am. Meteorol. Soc.*, *81*(9), 2035–2046, 2000.

Xie, P., and P. A. Arkin, A comparison of gauge observations and satellite estimates of monthly precipitation, *J. Appl. Meteor.*, *34*, 1143–1160, 1995.

Xie, P., and P. A. Arkin, Analysis of global monthly precipitation using gauge observations, satellite estimates and numerical model predictions, *J. Climate*, *9*, 840–858, 1996.

Xie, P., and P. A. Arkin, Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates and numerical model outputs, *Bull. Appl. Meteor.*, *34*, 1143–1160, 1997.

CHAPTER 24

RAINFALL

JAMES A. SMITH

The capability to measure rainfall advanced dramatically in the last quarter of the twentieth century. The advances have been paced by remote-sensing technologies including both ground-based weather radar and satellite-borne instruments. The most dramatic developments have centered around the capability to monitor precipitation globally from satellite sensors. This measurement capability provides a variety of avenues for hydroclimatological analysis and forecasting. Advances in ground-based radar technologies and deployment of dense networks of rain gages has enhanced the ability to measure rainfall at short time scales (less than 1 h) and small spatial scales (less than 1 km). These time and space scales are often most relevant for water management applications. A brief summary of rainfall measurement and analysis capabilities is presented in the following three sections and organized by the three principal measurement technologies: rain gage, radar, and satellite.

1 RAIN GAGES

Networks of rain gages play a key role in hydrologic applications ranging from flood forecasting to design of high-hazard structures and water supply management. A wide variety of recording and nonrecording rain gages are used for hydrologic applications. Review and discussion of rain gage technologies are presented in the work by Sumner (1988).

There exist several inherent sources of error that affect all types of rain gages. All rain gages suffer from errors due to modification of the wind field by the gage [see Robinson and Rodda (1969) for detailed discussions]. The magnitude of errors depends on wind speed, siting characteristics, and type of precipitation (Groisman

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

and Legates, 1994; Sevruk, 1982, 1989; Nystuen et al., 1996; Steiner et al., 1999; McCollum and Krajewski, 1998; Larson and Peck, 1974). Rain gage measurement of rainfall is especially difficult in a variety of settings, including mountain ridges, forests, and water bodies. Measurement errors for snow are typically much larger than for rain and are generally in the form of catch deficiencies (Groisman and Legates, 1994).

Rain gage networks serve as the basis for climatological assessments of precipitation that are used for a wide range of applications (see, e.g., Frei and Schaer, 1998). Three of the principal types of climatological analyses that are used for water management applications are illustrated in Figures 1 to 3. Assessments of average rainfall conditions, in a variety of forms, are central to activities involving industrial, municipal, and agricultural water use. Mean annual precipitation is shown in Figure 1 [see also Groisman and Legates (1994) for a discussion of biases in rain gage analyses of mean precipitation]. Global assessments of continental precipitation have been developed from rain gage observations by Legates (1987) [see also Legates and Wilmott (1990)]. Precipitation frequency analysis plays a central role in engineering design problems, especially in urban areas (Urbonas and Roesner, 1993). The 15-min, 100-year rainfall magnitude for the United States (Frederick et al., 1977) is illustrated in Figure 2. The network of gages that have the temporal resolution to provide short-term precipitation frequency analyses, such as those in Figure 2, is far less dense than the rain gage network used to produce mean annual precipitation maps. Consequently, it is difficult to assess the true geographic variability of extreme rainfall rates. It is likely that geographic features, such as mountains and land–water boundaries, exert a pronounced influence on the frequency of extreme rainfall rates. The density of the network, however, is not adequate to resolve these geographic variations. Design of high-hazard structures, such as spillways on major dams, is determined through probable maximum precipitation (PMP) analyses (Hansen, 1987; WMO, 1986). Rain gage data sets, in the form of storm catalogs, play a central role in PMP analyses. Storm catalogs for PMP analyses consist of gage observations from specific events. Consequently, the density of gage observations in regions experiencing catastrophic rainfall is critical for PMP analyses. The 6-h, 200 mi^2 PMP for the eastern United States is shown in Figure 3. The greatest uncertainties in PMP analyses are for small areas (less than 200 mi^2), short time periods (6 h and less), and for regions of complex terrain (National Research Council, 1994).

2 RADAR

Implementation of the NEXRAD (next-generation weather radar) system of WSR-88D (weather surveillance radar—1988 Doppler) radars has resulted in dramatic advances in rainfall measurement capabilities for the United States (Klazura and Imy, 1993). Operational National Weather Service (NWS) rainfall products derived from WSR-88D observations provide rainfall analyses for the United States at 1-h time resolution and spatial resolution of approximately 4 km (Hudlow et al., 1991).

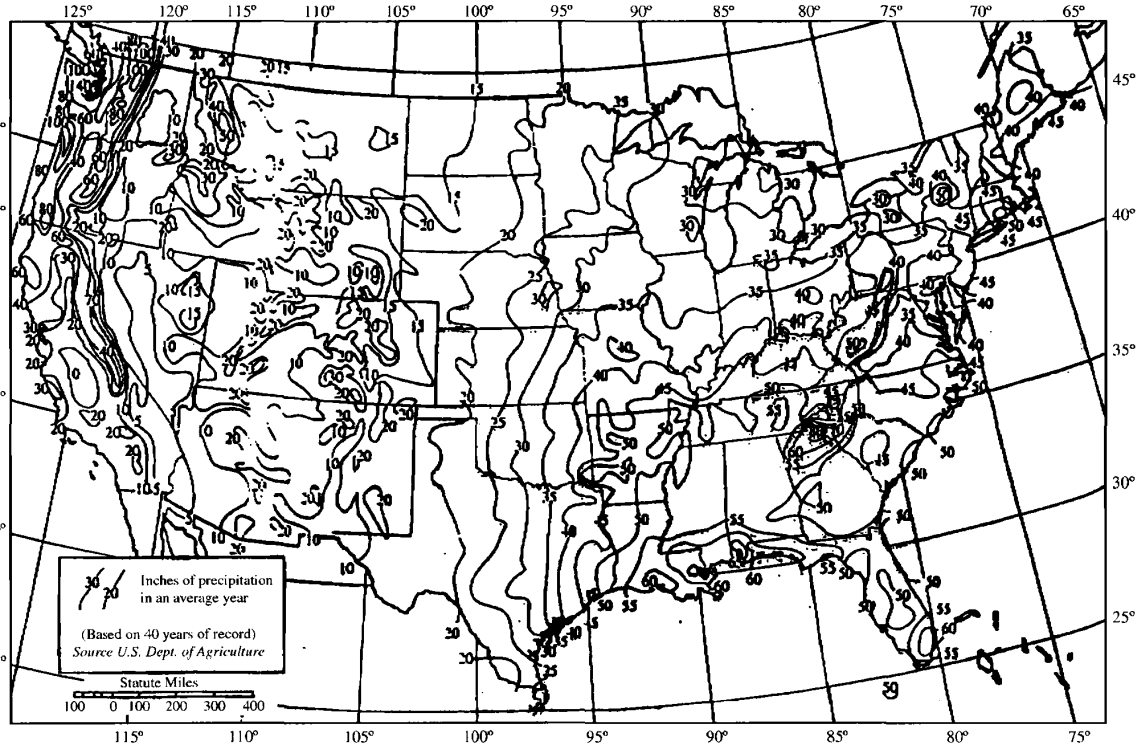


Figure 1 Mean annual precipitation (inches) for the United States from rain gage observations.

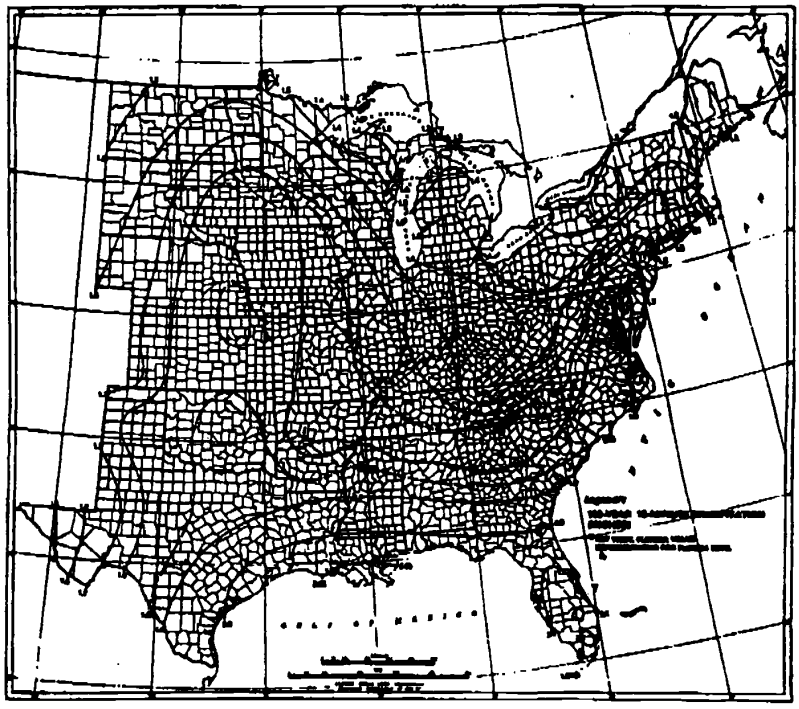


Figure 2 The 100-year, 15-min rainfall magnitudes (inches) for the United States east of the Rocky Mountains.

The hourly digital product (HDP) rainfall estimates are created by the WSR-88D radar product generator on a 131 × 131, 4-km grid centered at each radar site. The range over which rainfall products are constructed for each site is approximately 230 km. The algorithm used to construct this product (Fulton et al., 1998) consists of the following steps: (1) quality control, including identification and elimination of anomalous propagation returns, (2) conversion of radar reflectivity factor to rainfall rate through a Z-R relationship, (3) correction for range effects, (4) aggregation of rainfall estimates to hourly, 4-km grid scale, and (5) bias correction using rain gage observations. The HDP product is the base rainfall product from the NEXRAD system. Detailed assessments of HDP algorithm performance are presented in Smith et al. (1996b) and Baeck and Smith (1998) [see also Joss and Waldvogel (1989), Wilson and Brandes (1979), and Anagnostou and Krajewski (1998)].

In a second stage of WSR-88D rainfall processing, multisensor precipitation analyses employ rain gage observations and the 4-km HDP rainfall fields in an optimal estimation framework developed by Krajewski (1987) and Seo (1998a, 1998b). These rainfall fields are subsequently composited into a regional mosaic.

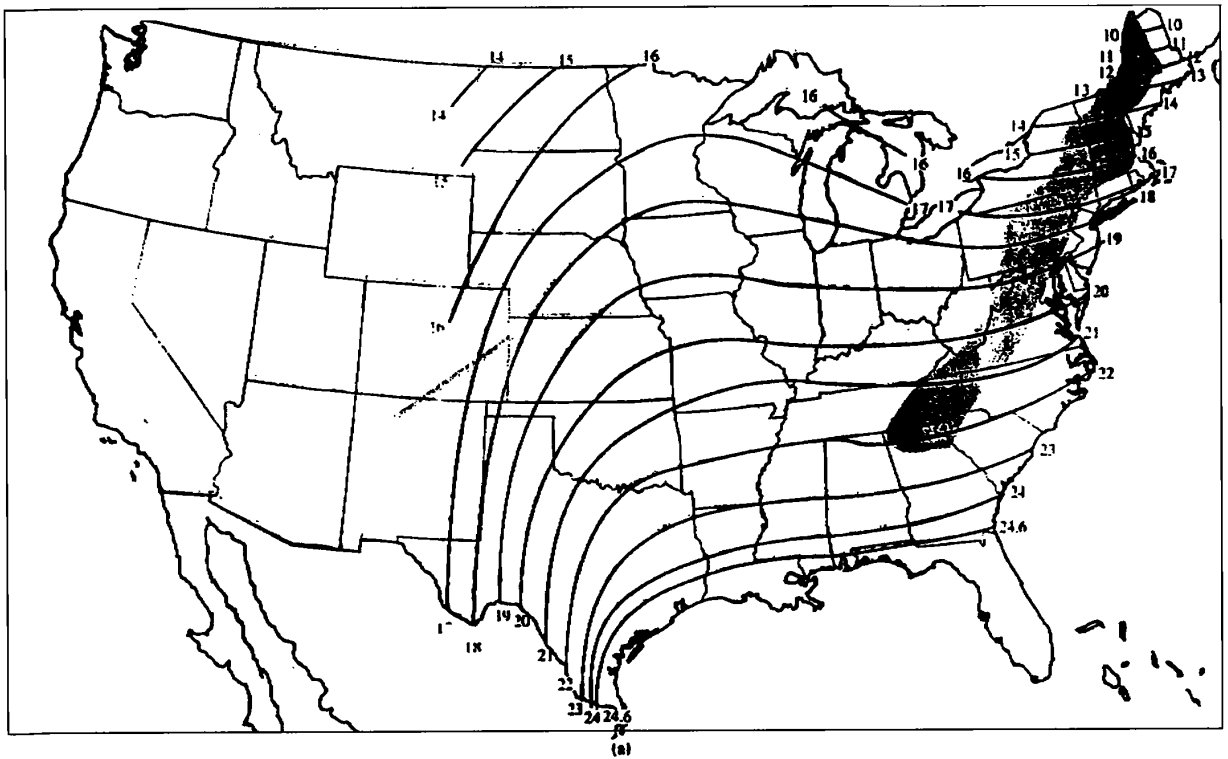


Figure 3 The 6-h, 200 mi^2 PMP magnitudes (inches) for the eastern United States.

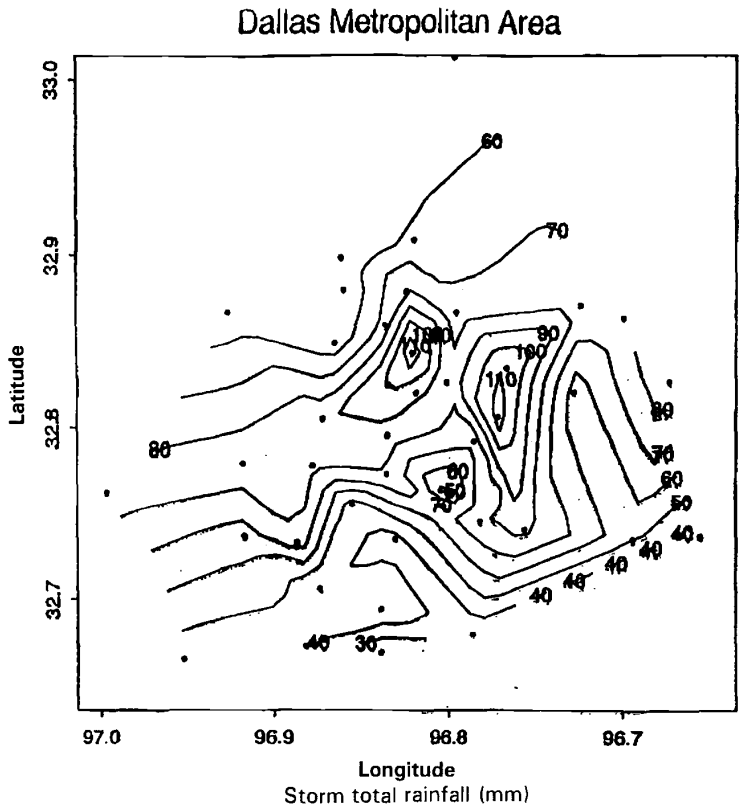


Figure 4 Storm total rainfall (mm) from the Dallas Metropolitan Area mesonet for the Dallas hailstorm of May 5, 1995. The dimensions of the surrounding box are approximately 30×30 km.

The regions that comprise the individual mosaics correspond to the watershed boundaries that delimit the NWS River Forecast Center areas of coverage. Algorithms used for mosaicking of multiple, overlapping radar coverages are described in Seo et al. (1998). A national, hourly precipitation analysis is produced at the National Centers for Environmental Prediction (NCEP).

The national 4-km, hourly rainfall mosaic produced by NWS from rain gage and WSR-88D rainfall products will provide an important source of rainfall information for climatological analyses, especially as the observing period increases. Radar observations have not generally served as the basis for climatological analyses of rainfall [see, however, Baeck and Smith (1995) for an exception]. Issues of bias in radar rainfall estimation must be addressed for radar-based rainfall databases to be most useful for climatological studies (Smith et al., 1996b).

Radar polarimetric measurements (Zrnic, 1996; Zrnic and Ryzhkov, 1996; Ryzhkov and Zrnic, 1996; Aydin et al., 1995), which utilize the capability of radar to transmit and receive electromagnetic radiation at alternating polarization, hold promise for providing significant improvements in rainfall estimates. Polarization measurements have been shown to be quite useful for quality control algorithms, including detection of bright band, hail, and AP [anomalous propagation (of radar waves, due to sharp gradients of water and air density)], as well as for algorithms for estimating rainfall rate (Peterson et al., 1999; Zrnic, 1996). The NEXRAD network was designed for eventual implementation of polarization measurements by the WSR-88D.

Radar has provided a significant component of the observational basis for studying storms that produce extreme rainfall. Chappell (1989) and Doswell et al. (1996) summarize key elements of heavy rainfall producing storms with particular emphasis on storms that produce large point rainfall accumulations through small net storm motion [see also Maddox et al. (1979)]. These storms have been termed quasi-stationary convective systems (Chappell, 1989). Houze et al. (1990) provide a detailed summary of radar-derived storm structure for severe thunderstorms in the central United States [see also Perica and Foufoula-Georgiou (1996) and Steiner et al. (1995)].

WSR-88D observations, and the rainfall products derived from these observations, have provided a new playing field for hydrologic application and science. Many hydrologic problems that were previously not possible to address due to an absence of information concerning rainfall, have been attacked from an observational perspective. Numerous examples can be drawn from flood hydrology. Figure 5 illustrates a storm total rainfall analysis constructed for the rapidan storm of June 27, 1995 (Smith et al., 1996a). More than 600 mm of rain fell on the east slope of the Virginia Blue Ridge during a 12-h period resulting in record unit discharge for the United States east of the Mississippi River and catastrophic landslides and debris flows. Fluvial and geomorphic impacts of the rapidan storm rival those described in the classic study by Hack and Goodlett (1960) for the June 1949 storm in the Shenandoah Mountains. The chief difference between studies of the 1949 and 1995 storms is rainfall measurement at the 1-km horizontal scale and 6-min time scale for the 1995 storm that allows direct assessment of hydrologic processes.

3 SATELLITE

Satellite-borne instruments have proven useful for monitoring precipitating cloud system since the 1960s. Steady progress has been made in developing algorithms for retrieving rainfall accumulations from passive satellite observations in the microwave (Negri et al., 1994; Adler et al., 1994) and infrared (Vicente and Scofield, 1997; Huffman et al., 1995; Adler and Negri, 1988) portions of the electromagnetic spectrum. This progress is reflected in rapidly advancing capabilities for hydroclimatological analysis (Kummerow et al., 2000; Adler et al., 2000; Huffman et

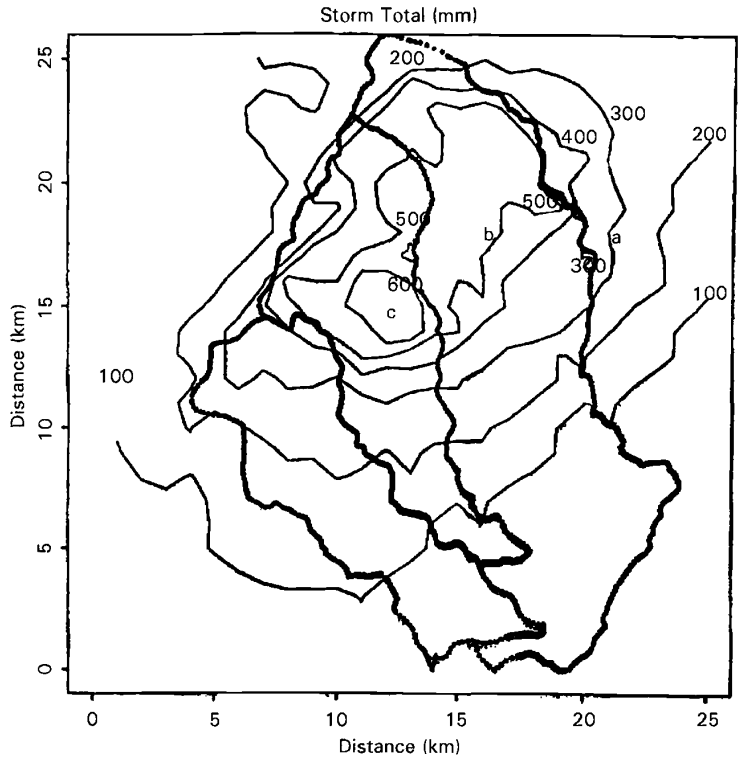


Figure 5 Storm total rainfall (mm) from the Sterling, Virginia, WSR-88D for the rapidan storm of June 27, 1995. The basin boundary for the 295-km^2 watershed and boundaries for 3 subwatersheds are shown in solid lines.

al., 1997, 2001; Krajewski et al., 2000). Techniques for quantitative precipitation estimation from satellite sensors are reviewed below.

A geostationary, infrared-based satellite algorithm (Vicente and Scofield, 1997) has been developed and implemented for heavy rainfall measurement. This algorithm is specifically designed for deep, moist convective systems. Estimated precipitation rates are based on the cloud top temperature obtained from the 10.7-μm infrared channel. The empirical equations used to relate cloud top temperature and rainfall rate were calibrated from radar data sets consisting of observations from thunderstorm systems. A moisture correction factor obtained from the precipitable water and mean relative humidity is used to adjust the estimates for different moist environments. The technique of relating rain rate and cloud top temperature tends to overestimate the rain area in some cases and rain rate in others. The technique is also

subject to underestimation of rain rates in warm cloud top environments and overestimation of cold top storms in strong wind shear environments.

The *Tropical Rainfall Measuring Mission* (*TRMM*) satellite (Simpson et al., 1988) is designed to measure tropical precipitation and its variation. With the inclusion of a precipitation radar, TRMM provides the first opportunity to estimate the vertical profile of the latent heat that is released through condensation. The TRMM rainfall data will be particularly important for studies of the global hydrological cycle and for testing the ability of climate models to simulate climate accurately on the seasonal time scale.

The TRMM instruments for rainfall observation consist of a precipitation radar, a multifrequency microwave radiometer, and a visible and infrared (VIS/IR) radiometer. The precipitation radar provides measurements of the three-dimensional rainfall distribution over both land and ocean. The precipitation radar will permit the measurement of rain over land where passive microwave channels have difficulty. The horizontal resolution is approximately 4 km, the range resolution is 250 m, and the scanning swath width is 220 km. The multichannel microwave radiometer provides information on vertically integrated precipitation, its areal distribution, and its intensity. Rainfall analyses using the microwave radiometer are best suited for open ocean conditions. The visible infrared (IR) scanner provides high-resolution information on cloud coverage, type, and cloud top temperatures and serves as the link between these data and the long and virtually continuous coverage by the geosynchronous meteorological satellites. The instrument, with a swath width of 720 km, will provide cloud distributions by type and height and rain estimates from brightness temperatures at a horizontal resolution of approximately 2 km.

Satellite IR observations from geostationary satellites have been used extensively for assessing the climatology of extreme rainfall producing storms. An extensive climatology has been developed for mesoscale convective complexes (Maddox, 1980) based on IR-based assessments of cloud properties. Numerous studies have examined the links between mesoscale convective complexes (MCCs), and the more general category of mesoscale convective systems, and heavy rainfall [see Houze (1993)].

REFERENCES

Adler, R. F., and A. J. Negri, A satellite infrared technique to estimate tropical convective and stratiform rainfall, *J. Appl. Meteor.*, *27*, 30–51, 1988.

Adler, R. F., G. J. Huffman, and P. R. Keehn, Global tropical rain estimates from microwave-adjusted geosynchronous IR data, *Remote Sensing Rev.*, *11*, 125–152, 1994.

Adler, R. F., G. J. Huffman, D. T. Bolvin, S. Curtis, and E. J. Nelkin, Tropical rainfall distributions determined using TRMM combined with other satellite and rain gauge information, *J. Appl. Meteor.*, *39*(12), 2007–2023, 2000.

Anagnostou, E., and W. F. Krajewski, Calibration of the WSR-88D precipitation processing

Aydin, K., V. N. Bringi, and L. Liu, Rain rate estimation in the presence of hail using S-band specific differential phase and other radar parameters, *J. Appl. Meteor.*, *34*, 404–410, 1995.

Baeck, M. L., and J. A. Smith, Climatological analysis of manually digitized radar data for the United States, *Water Resour. Res.*, *31*(12), 3033–3049, 1995.

Baeck, M. L., and J. A. Smith, Rainfall estimation by the WSR-88D for heavy rainfall events, *Weather Forecast.*, *13*, 413–436, 1998.

Barros, A. P., and D. P. Lettenmaier, Dynamic modeling of orographically induced precipitation, *Rev. Geophys.*, *32*(3), 265–284, 1994.

Chappell, C., Quasistationary convective events, in P. Ray (Ed.), *Mesoscale Meteorology*, American Meteorological Society, Boston, 1989.

Doswell III, C. A., H. E. Brooks, and R. A. Maddox, Flash flood forecasting: An ingredients-based methodology, *Weather Forecast.*, *11*(4), 560–581, 1996.

Frederick, R. H., V. A. Myers, and E. P. Auciello, Five- to 60-minute precipitation frequency for the eastern and central United States, NOAA Technical Memo, NWS Hydro-35, June 1977.

Frei, C., and C. Schaer, A precipitation climatology of the Alps from high-resolution rain-gauge observations, *Int. J. Climatol.*, *18*(8), 873–900, 1998.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, and D. A. Miller, The WSR-88D rainfall algorithm, *Weather Forecast.*, *13*(2), 377–395, 1998.

Groisman, P. Y., and D. R. Legates, The accuracy of United States precipitation data, *Bull. Am. Meteor. Soc.*, *75*(2), 215–227, 1994.

Hack, J., and J. Goodlett, Geomorphology and forest ecology of a mountain region in the central Appalachians, *U.S. Geological Survey Professional Paper 347*, 1960.

Hansen, M., Probable maximum precipitation for design floods in the United States, *J. Hydrol.*, *96*, 267–278, 1987.

Houze, R. A., *Cloud Dynamics*, Academic, New York, 1993.

Houze, R. A., B. F. Smull, and P. Dodge, Mesoscale organization of springtime rainstorms in Oklahoma, *Monthly Weather Rev.*, *118*, 613–654, 1990.

Hudlow, M. D., J. A. Smith, M. L. Walton, and R. C. Shedd, NEXRAD—New era in hydrometeorology, in I. Cluckie and C. Collier (Eds.), *Hydrological Applications of Weather Radar*, Ellis-Norwood, New York, 1991, pp. 602–612.

Huffman, G. J., R. F. Adler, B. Rudolf, U. Schneider, and P. R. Keehn, A technique for combining satellite-based estimates raingauge analysis, and NWP model precipitation information into global precipitation estimates, *J. Climate*, *8*(5), 1284–1295, 1995.

Huffman, G. J., R. F. Adler, P. Arkin, A. Chang, R. Ferraro, A. Gruber, J. Janowiak, A. McNab, B. Rudolf, and U. Schneider, The Global Precipitation Climatology Project (GPCP) combined precipitation dataset, *Bull. Am. Meteor. Soc.*, *78*(1), 5–20, 1997.

Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, Global precipitation at one-degree daily resolution from multisatellite observations, *J. Hydrometeor.*, *2*(1), 36–50, 2001.

Joss, J., and A. Waldvogel, Precipitation measurement and hydrology: A review, in *Battan Memorial and Radar Conference*, David Atlas, editor, American Meteorological Society, Boston, 1989.

Klazura, G. E., and D. A. Imy, A description of the initial set of analysis products available from the NEXRAD WSR-88D System, *Bull. Am. Meteor. Soc.*, *74*, 1293–1311, 1993.

Krajewski, W. F., Co-kriging of radar and rain gage data, *J. Geophys. Res.*, *92*(D8), 9571–9580, 1987.

Krajewski, W. F., G. J. Ciach, and J. R. McCollum, Initial validation of the global precipitation climatology project monthly rainfall over the United States, *J. Appl. Meteorol.*, *39*(7), 1071–1086, 2000.

Kummerow, C., J. Simpson, O. Thiele, W. Barnes, A. T. C. Chang, E. Stocker, R. F. Adler, A. Hou, R. Kakar, F. Wentz, P. Ashcroft, T. Kozu, Y. Hong, K. Okamoto, T. Iguchi, H. Kuroiwa, E. Im, Z. Haddad, G. Huffman, B. Ferrier, W. S. Olson, E. Zipser, E. A. Smith, T. T. Wilheit, and G. North, The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit, *J. Appl. Meteor.*, *39*(12), 1965–1982, 2000.

Larson, L., and E. Peck, Accuracy of precipitation measurements for hydrologic modeling, *Water Resour. Res.*, *10*(4), 857–863, 1974.

Legates, D. R., A climatology of global precipitation, *Climatology*, *40*(1), 1987.

Legates, D. R., and C. J. Willmott, Mean Seasonal and Spatial Variability in Gauge-Corrected, Global Precipitation, *Int. J. Climatol.*, *10*(2), 111–127, 1990.

Maddox, R. A., Mesoscale convective complexes, *Bull. Am. Meteor. Soc.*, *61*(11), 1374–1387, 1980.

Maddox, R., C. Chappell, and L. Hoxit, Synoptic and meso-α scale aspects of flash flood events, *Bull. Am. Meteor. Soc.*, *60*(2), 115–123, 1979.

McCollum, J. R., and W. F. Krajewski, Uncertainty of monthly rainfall estimates from rain gauges in the Global Precipitation Climatology Project, *Water Resour. Res.*, *34*(10), 2647–2654, 1998.

National Research Council, *Estimating Bounds on Extreme Precipitation Events*, National Academy Press, Washington DC, 1994.

Negri, A. J., R. F. Adler, E. J. Nelkin, and G. J. Huffman, Regional rainfall climatologies derived from special sensor microwave imager (SSM/I) data, *Bull. Am. Meteor. Soc.*, *75*, 1165–1182, 1994.

Nystuen, J. A., J. R. Proni, P. G. Black, and J. C. Wilkerson, A comparison of automatic rain gauges, *J. Atmos. Oceanic Technol.*, *13*, 62–73, 1996.

Perica, S., and E. Foufoula-Georgiou, Linkage of scaling and thermodynamic parameters of rainfall: Results from midlatitude mesoscale convective systems, *J. Geophys. Res.*, *101*(D3), 7431–7448, 1996.

Petersen, W. A., L. D. Carey, S. A. Rutledge, J. C. Knievel, N. J. Doesken, R. H. Johnson, T. B. McKee, T. Vonder Haar, and J. F. Weaver, Mesoscale and radar observations of the Fort Collins flash flood of 28 July 1997, *Bull. Am. Meteor. Soc.*, *80*(2), 191–216, 1999.

Robinson, A. C., and J. C. Rodda, Rain, wind and the aerodynamic characteristics of raingauges, *Met. Mat.*, *98*, 113–120, 1969.

Ryzhkov, A. V., and D. S. Zrnic, Assessment of rainfall measurement that uses specific differential phase, *J. Appl. Meteorol.*, *35*(11), 2080–2090, 1996.

Seo, D.-J., Optimal estimation of rainfall fields using radar and rain gage data, *J. Hydrol.*, *208*(1,2), pp. 25–36, 1998.

Seo, D.-J., Real-time estimation of rainfall fields using rain gage data under fractional coverage conditions, *J. Hydrol.*, *208*(1,2), pp. 37–52, 1998.

Seo, D. J., J. Breidenbach, and E. Johnson, Real-time estimation of mean field bias in radar rainfall data, *J. Hydrol.*, *223*(3,4), pp. 131–147, 1999.

Seo, D. J., R. Fulton, and J. Breidenbach, Rainfall estimation in the WSR-88D era for operational hydrologic forecasting in the National Weather Service, Symposium on Hydrology, Phoenix, American Meteorological Society, 1998, pp. J60–J62.

Sevruk, B., Methods of correction for systematic error in point precipitation measurement for operational use, WMO Publication No. 589, OHR No. 21, 1982.

Sevruk, B., Wind-induced measurement error for high intensity rains, Proceedings, WMO/IAHS International Workshop on Precipitation Measurements, St. Moritz, Switzerland, 1989, pp. 199–204.

Simpson, J., R. F. Adler, and G. North, A proposed tropical rainfall measuring mission (TRMM), *Bull. Am. Meteor. Soc.*, *69*, 278–295, 1988.

Smith, J. A., M. L. Baeck, M. Steiner, and A. J. Miller, Catastrophic rainfall from an upslope thunderstorm in the Central Appalachians: the Rapidan Storm of June 27, 1995, *Water Resour. Res.*, *32*(10), 3099–3113, 1996a.

Smith, J. A., D.-J. Seo, M. L. Baeck, and M. D. Hudlow, An intercomparison study of NEXRAD precipitation estimates, *Water Resour. Res.*, *32*(7), 2035–2045, 1996b.

Steiner, M., R. A. Houze, Jr., and S. E. Yuter, Climatological characterization of three-dimensional storm structure from operational radar and raingage data, *J. Appl. Meteor.*, *34*, 1978–2007, 1995.

Steiner, M., J. A. Smith, S. J. Burges, C. V. Alonso, and R. W. Darden, Effect of bias adjustment and rain gauge data quality control on radar rainfall estimation, *Water Resour. Res.*, *35*(8), 2487–2503, 1999.

Sumner, G., Precipitation measurement and observation, in *Precipitation: Process and Analysis*, Wiley, Chichester, 1988, Chapter 7.

Vicente, G. A., and R. A. Scofield, Real time rainfall rate estimates derived from the GOES-8/9 satellites for flash flood forecasting, numerical modeling and operational hydrology, Preprints, 13th Conference on Hydrology, Long Beach, CA, American Meteorological Society, 1997, pp. J115–J118.

Urbonas, B. R., and L. A. Roesner, Hydrologic design for urban drainage and flood control, in D. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1992.

World Meteorological Organization (WMO), *Manual for estimation of Probable Maximum Precipitation*, 2nd ed., WMO-No. 332, WMO, Geneva, Switzerland, 1986.

Wilson, J. W., and E. A. Brandes, Radar measurement of rainfall—a summary, *Bull. Am. Meteor. Soc.*, *60*, 1048–1058, 1979.

Zrnic, D. S., Weather radar polarimetry—trends toward operational applications, *Bull. Am. Meteorol. Soc.*, *77*(7), 1529–1534, 1996.

Zrnic, D. S., and A. V. Ryzhkov, Advantages of rain measurements using specific differential phase, *J. Atmos. Ocean. Tech.*, *13*(2), 454–464, 1996.

CHAPTER 25

SNOW HYDROLOGY AND WATER RESOURCES (WESTERN UNITED STATES)

ROGER C. BALES AND DON CLINE

1 INTRODUCTION

Seasonally snow-covered areas of Earth offer special challenges for water resources management, challenges that arise from both hydrologic and social factors. Seasonal snowpacks account for the major source of the runoff for streamflow and groundwater recharge over wide areas of the midlatitudes. For example, in the western United States over 85% of the annual runoff from the Colorado River basin originates as snowmelt. Most of this is from a few small source areas in four western states, mostly above 2700 m, which comprise only 12% of the basin area. Globally, snowmelt runoff from Earth's mountains fills the rivers and recharges the aquifers that over a billion people depend on for their water resources. Future climate variability and change are expected to result in major changes in the partitioning of snow and rainfall and the timing of snowmelt, which will have important implications for water use and resource management in these regions. It is therefore important to understand the processes controlling snowmelt runoff for both water resources as well as other resource management purposes.

2 CURRENT HYDROCLIMATIC CONDITIONS IN THE WESTERN UNITED STATES

Much of the variability in snow cover found in Earth's seasonally snow-covered regions can be found in the relatively well-studied western United States, which

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

has regions ranging from the high-precipitation Pacific Northwest coast to the semi-arid Southwest. The area from the Rocky Mountains to the Pacific Coast can be conveniently divided into seven regions with different hydroclimatic regimes (Fig. 1). Throughout the region, much of the annual streamflow is directly attributable to springtime melting of snow accumulation from the previous winter; however, there are also lower-elevation areas within the region that experience snowmelt throughout the winter and spring. Extended winter snowpack water storage in alpine areas, with often gradual melt rates, results in annual hydrographs having rising limbs of characteristically low slope, usually superimposed with small diel fluctuations reflecting daily melt cycles (Fig. 2). Beyond this basic similarity, however, wide differences in the source, delivery, and amount of moisture each region receives, the amount of water typically stored in their snowpacks, and the rate of release of that water result in different streamflow regimes between regions.

The climate of the western United States is dominated by large-scale atmospheric circulation originating over the north Pacific. In winter, the Pacific/North American (PNA) anomaly pattern forms a series of pressure centers of alternating sign stretching across the Pacific into southern Canada and down toward the Gulf of Mexico (Lin et al., 1990). The PNA typically forms a series of dry longwave ridges and wet troughs across North America with cold polar air masses to the north of the frontal boundary and warm subtropical air masses to the south. The position of the ridges and troughs influences the seasonal moisture cycle. The mountain ranges of the western United States, like all major mountain ranges, strongly influence global atmospheric circulation; in doing so, they affect the seasonal moisture cycle.

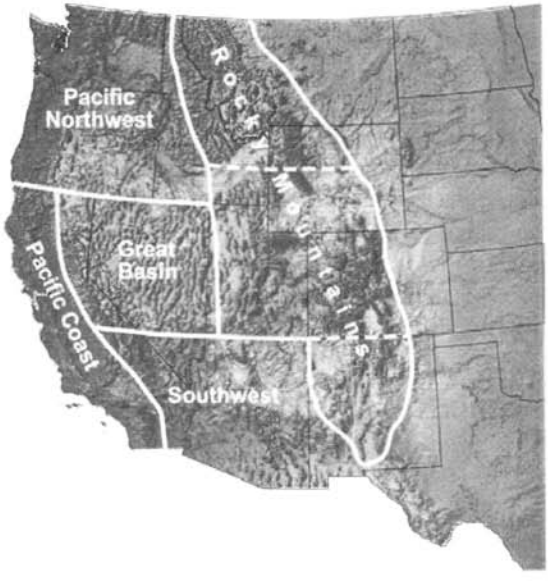


Figure 1 Seven hydroclimatic regimes of the western United States [after Paulson et al. (1991)].

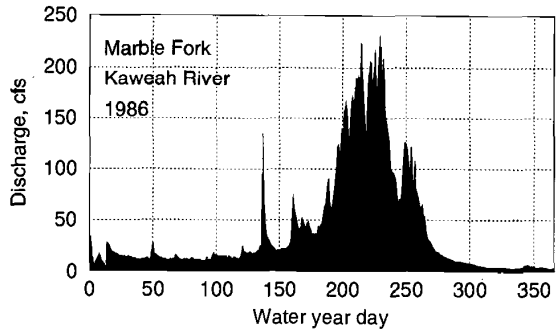


Figure 2 Discharge hydrograph for 19 km^2 Marble Fork of Kaweah River, southern Sierra.

The major source of moisture for all of the western United States is the Pacific Ocean; during fall and winter, orographic lifting and cooling of Pacific air masses laden with moisture results in precipitation either as rain or snow. The coast ranges, Cascades and Sierra Nevada, form a major orographic barrier for the Pacific moisture, causing much of the winter precipitation to fall as rain on the western side of the mountains. Winter precipitation on the eastern side of the Cascades and Sierra Nevada, although less, generally falls as snow in higher elevations. Relatively warm winter temperatures usually result in warm, wet snowpacks that often are nearly isothermal, and susceptible to rapid melting from warm temperatures and rainfall.

In the spring and summer, moisture from the Gulf of Mexico and subtropical Atlantic Ocean becomes important in most of the western states, with the exception of the coastal states. Early spring Gulf and subtropical Atlantic moisture often precipitates as snow, especially at higher elevations. Additional summertime moisture is provided in the southwestern states by subtropical Pacific air masses. With the exception of "land-recycled" moisture from land surface evapotranspiration (Paulson, 1991), these are the three sources of moisture that provide the western United States with precipitation and runoff. Peak snow accumulation and snowpack water storage in most of the region is found at higher elevations and generally occurs in March or April with snowmelt runoff occurring through May to July, depending on elevation and latitude.

Frontal activity associated with low-pressure systems is responsible for much of the winter precipitation in the northern Rocky Mountains, and upslope transport of moisture from east to west is important at lower elevations on the eastern side of the mountains. Summer precipitation, much of which ends up as evapotranspiration in the semiarid western United States, is mostly influenced by convective activity. However, snowpack storage serves as the major water supply for the summer months. The continentality of the northern Rocky Mountain region leads to cold, dry snowpacks. Significant energy is required to raise the temperature of the snowpack to the isothermal and melting stage; as a result the snowpack tends to remain

well into spring. Rainfall generally does not contribute sufficient energy to drive snowmelt, until perhaps very late in the season.

High elevations in the central Rocky Mountains receive most of this region's annual precipitation as winter snowfall. Pacific frontal systems bringing most of the winter moisture to this region can arrive from the west, northwest, or southwest, and this influences the distribution of precipitation. Westerly tracks are orographically lifted to some extent by the Wasatch Plateau in Utah and are lifted further by the ranges along the Continental Divide in central Colorado, resulting in the heaviest precipitation west of the Continental Divide. Northwesterly tracks are lifted by the Wasatch Range and the Uinta Mountains in Utah and by the ranges along the Divide in north central Colorado, resulting in heavier precipitation at these locations. Storm tracks arriving from the southwest do not encounter major orographic effects until they reach the San Juan Mountains in southwestern Colorado, resulting in typically heavy winter precipitation in this part of the region from these storm tracks. In general, precipitation declines markedly throughout areas east of the Continental Divide. However, low-pressure systems east of the Divide can bring significant moisture in from the Gulf of Mexico during spring, resulting in sometimes heavy snowfall in the foothills at lower elevations on the eastern side of the Divide. Lower elevation areas of the central Rockies receive considerably less precipitation; most of the region's snowpack storage is concentrated in the higher mountains.

3 MEASUREMENT AND ESTIMATION OF SNOW PROPERTIES

Historical Background

Undoubtedly, the main recurring question in snow hydrology in the western United States is: How much snow is out there? Water resources managers forecast the amount of seasonal runoff, based in part on estimates of the amount of snow accumulation, or snow water equivalent (SWE), across a watershed or region and in part on forecasts of future precipitation. Estimates of SWE and snow-covered area (SCA) are used for a variety of purposes that are vital to the economy of a region, including: reservoir management, snow load maps, annual precipitation maps (for planning), drought monitoring, fish and game management, recreation (e.g., skiing, river trips), acid precipitation monitoring, and avalanche forecasting. (See Section 5 by Doesken.)

Historically, the Natural Resource Conservation Service (NRCS) has been charged with coordinating snow surveys, or point measurements of SWE. It also prepares seasonal water supply outlooks in the western United States. Predictions of water availability in the western United States are made by inventorying snowpacks in winter and early spring using measurements at over 2000 snow courses, including about 1000 snowpack telemetry (SNOTEL) sites that provide continuous data. The remaining sites are manual and are visited monthly. Empirical relationships between these observations and measured streamflow are used to forecast streamflow at over 500 points. In California, the California Cooperative Snow Survey (CCSS) coordi-

nates measurements; it depends on 40 cooperating agencies for data collection. CCSS makes seasonal water supply forecasts, as do many program cooperators; weekly updates are made for major streams (Hart and Gehrke, 1990).

Estimation of the spatial distribution of SWE is challenging because of the many factors that affect its distribution and the small correlation length of the SWE spatial distribution. Topographic heterogeneity and variability in precipitation patterns also present problems in accurately determining the time of maximum accumulation. The simplicity of regression models makes them an attractive means of estimating SWE because of the large amount of work required to directly measure SWE on the catchment scale.

Estimates of the spatial distribution of SWE for the western United States come primarily from two sources. Operationally, the National Weather Service's (NWS's) National Operational Hydrologic Remote Sensing Center (NOHRSC) assimilates in situ and airborne snow survey data with satellite observations of snow cover to estimate SWE distributions throughout the winter and spring (Figs. 3*a* and 3*b*). Second, atmospheric models estimate the distribution of SWE based on modeled snowfall and surface energy balance parameters. The NWS SWE estimates are based on an interpolation procedure called snow estimation and updating system (SEUS), which interpolates between observed points to produce gridded SWE estimates. Point SWE estimates come from the NRCS sites and remote sensing. The NOHRSC conducts airborne SWE surveys along 1800 flight lines throughout the United States, many of which are located in the west (Fig. 4). Water contained in the snowpack attenuates terrestrial gamma radiation. The attenuation is measured relative to snow-free conditions to estimate SWE.

Remote Sensing

Remote sensing provides important spatial information about snow that can be used to improve the accuracy and timeliness of hydrologic forecasts for seasonally snow-covered areas, with commensurate gains in water resources management. At present, the only remotely sensed snow information used in operational hydrologic forecasting is the areal extent of snow cover (Fig. 3*a*). Over the past decade there has been an expanded development of remote sensing as a tool for determining other snow properties, which can be used to assist in estimating snow distributions and snowmelt runoff. There has also been a move toward development of physically based snowmelt models to use with this emerging data, particularly for alpine areas. The coupling of remote sensing and physically based approaches will enable making not only more accurate basin-scale forecasts but will also provide spatially distributed estimates of snowmelt.

The possibilities for detecting snowpack properties are largely determined by the wavelength being recorded by the remote-sensing instrument. Visible and near-infrared wavelengths, because they do not penetrate far into the snowpack, mainly provide information about the surface of the snowpack (e.g., snow-covered area, grain size, and albedo). However, microwave wavelengths can penetrate the snow-

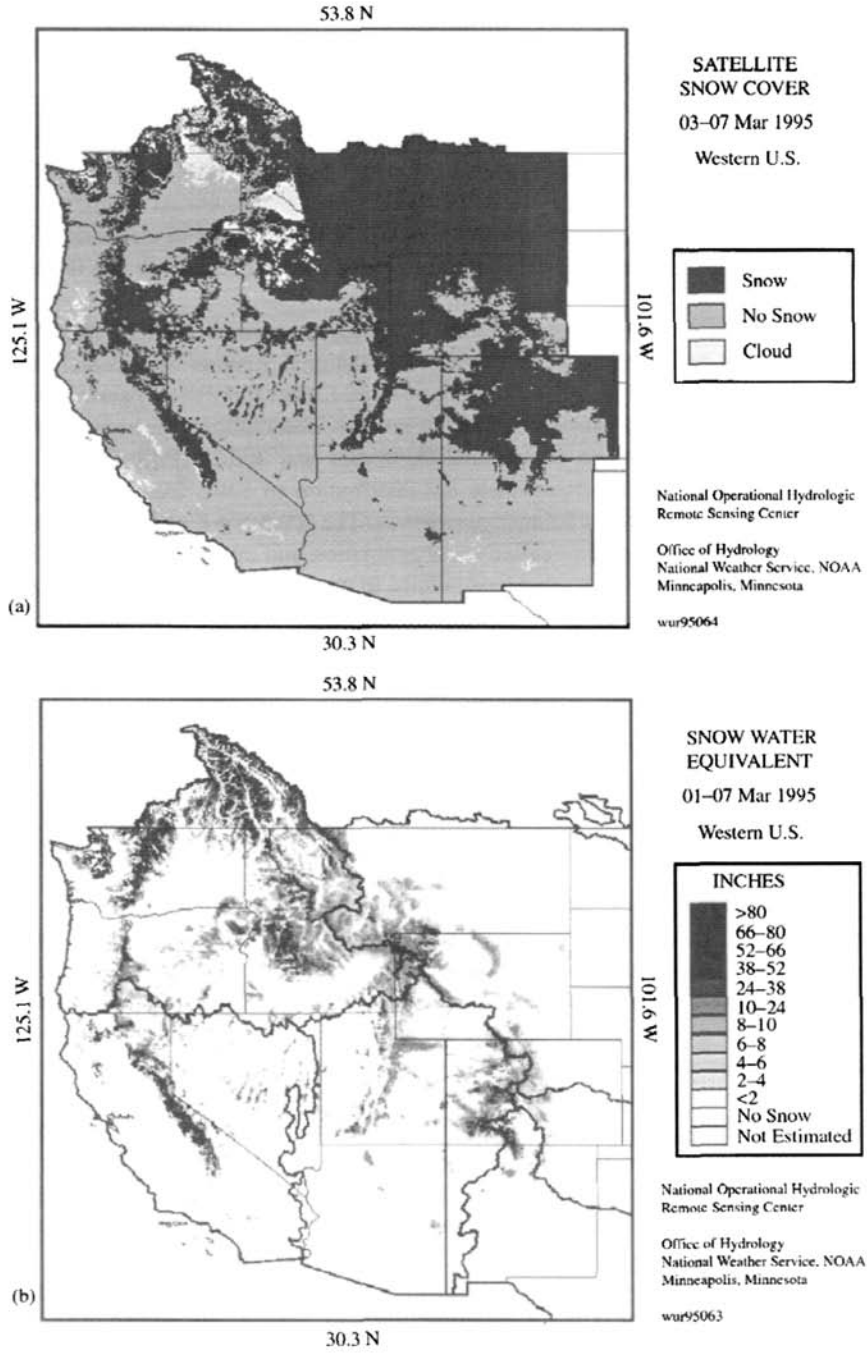


Figure 3 Operational snow products from the National Weather Service: (*a*) Satellite snow cover for March 1995 and (*b*) snow water equivalence for March 1995.

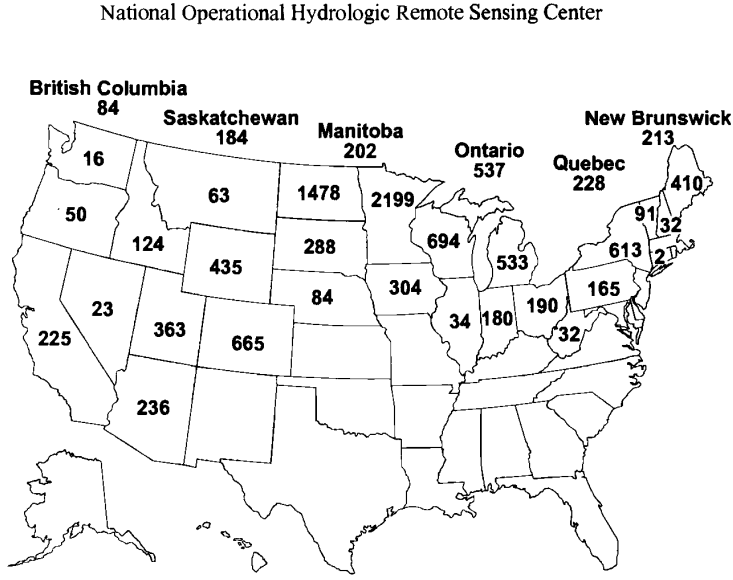


Figure 4 Number of airborne snow survey flight lines by state (1980–1998). National Operational Hydrologic Remote Sensing Center (NOHRSC).

pack, thereby providing an opportunity to collect volume integrated data (e.g., SWE).

Because of the difficulty of making field measurements in snow-covered mountainous regions, remote sensing has been pursued as a means of measuring snow-cover properties. The National Oceanic and Atmospheric Administration's (NOAA's) advanced very-high resolution radiometer (AVHRR) data have been routinely used for classification of snow-covered versus snow-free area (Matson et al., 1986; Matson, 1991; Xu et al., 1993). Differences between the spatial, temporal, spectral, and radiometric resolutions of different remote-sensing instruments result in trade-offs between instruments for hydrologic applications. Optimization of one type of resolution generally involves some sacrifice in other types of resolution. For example, the *Landsat* thematic mapper (TM) has a much better spatial resolution than the AVHRR (30 m versus 1 km pixel size, respectively); however, the AVHRR can provide daily coverage of a given point, whereas the TM can only provide biweekly coverage. Development of accurate snow-cover information for areas with steep, variable topography characteristic of the western United States requires higher resolution data than are currently available from operational remote-sensing instruments or improved processing of the current data.

Passive microwave sensors are used to monitor snow, but three problems have limited its application. First, uncertainty in snow texture results in a significant noise in the calibration of a brightness temperature index with measured snow properties. Second, passive microwave imagery has a large pixel size that results in significant mixed pixels over areas with forest, mountains, or lakes. It appears unlikely that snow properties could be "unmixed" in these situations. This restricts operational use to large flat areas such as prairies and tundra. Finally, the signature from snow is indistinguishable from bare ground when the snow is wet, which requires special processing of time series and inference.

For mapping of snow properties of greatest hydrological importance, a synthetic aperture radar (SAR) with some special characteristics is necessary. A SAR is sensitive to many snow parameters such as snow density, depth grain size, free-liquid water content, and snow-pack structures that hydrologists use. It can image day or night in all weather, and it has a fine spatial resolution compatible with topographic variation affecting snow distribution. Experiments as part of the SIR-C/X SAR missions have significantly advanced and demonstrated the capabilities of new multi-frequency and multipolarization SAR—to map both wet and dry snow covers (Shi and Dozier, 1997), to infer snow wetness (Shi and Dozier, 1995), and to estimate snow density and depth and thereby snow water equivalent (Shi and Dozier, 1996).

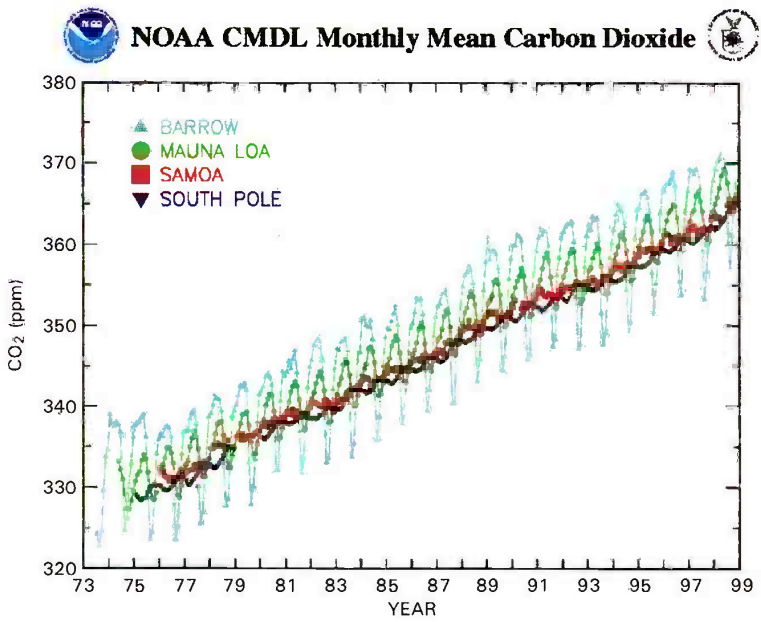
However, operational satellites do not provide the necessary data. To achieve sufficiently high spatial resolution to measure the variability of SWE in mountainous areas, a multiple-frequency, multiple-polarization SAR is required. At present, the measurement of the spatial distribution of SWE, and total snow volume within a mountainous basin, must be performed by intensive field sampling to attempt to represent the large spatial variability of alpine snowpacks. Logistics and safety limitations generally restrict the number of field samples that may be so obtained (Elder et al., 1991a). Thus the problem of determining the volume and distribution of snowpack water storage within mountain basins remains acute.

Climate Models

While still not an operational tool, the use of high-resolution regional climate models for simulating seasonal and interannual changes in snowpack in areas such as the western United States is quite promising. At a 60-km resolution such a model can reproduce the overall patterns of measured precipitation and snow cover and to some extent year-to-year variations (Figs. 5*a* and 5*b*) (Seth et al., 1999). Errors in simulating the actual magnitude of seasonal snow accumulation arise in large part from lack of realism in model topography, with inadequacies in model parameterization also a factor. However, climate modeling offers great promise as a tool that can be integrated with ground-based and satellite observations.

Emerging Technologies

The National Aeronautic and Space Administration's (NASA's) moderate resolution imaging spectrometer (MODIS) will provide near-daily global coverage (comparable



Atmospheric carbon dioxide mixing ratios determined from the continuous monitoring programs at the 4 NOAA CMDL baseline observatories. Principal investigator: Pieter Tans, NOAA CMDL Carbon Cycle Group, Boulder, Colorado, (303) 497-6678. ptans@cmdl.noaa.gov.

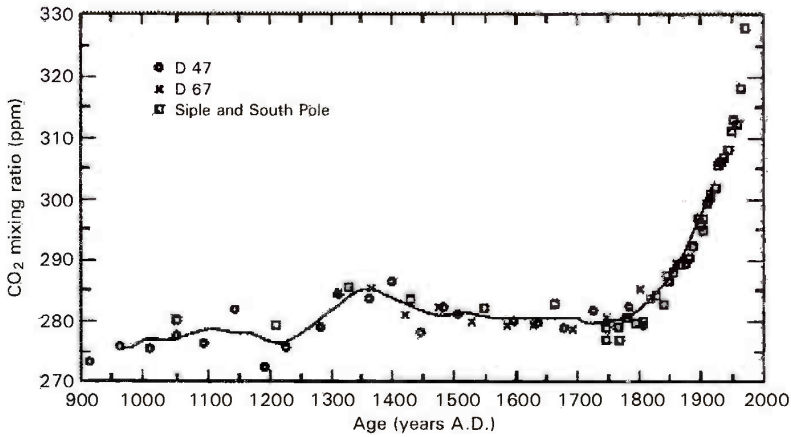


Figure 5 (Chapter 1) (*a*) Monthly concentrations of CO_2 measured from gas samples at four monitoring sites operated by NOAA's Climate Monitoring and Diagnostics Laboratory from the early 1970s; (*b*) CO_2 concentrations determined from ice core samples estimated to go back ~1000 years.

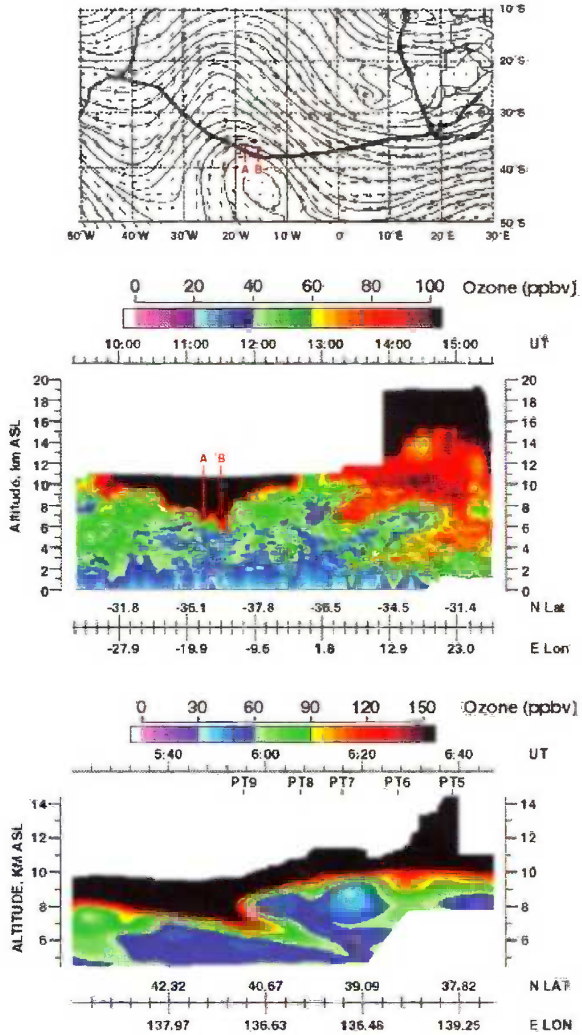


Figure 10 (Chapter 1) Three-panel figure showing evidence of ozone input from the stratosphere into the troposphere in both hemispheres. The top panel shows the flight path (heavy line) of a DC-8 airplane on October 3, 1992, from South America to Africa that intersected a trough protruding from higher latitudes. Points A and B on that flight path show high concentrations of ozone being transported to altitudes below 6 km in the middle panel; the data depicted in this panel were obtained from a differential absorption lidar system that measured ozone below the 11-km flight level of the DC-8. The lowest panel shows a similar feature for a flight on March 11, 1994, in the Northern Hemisphere. As the airplane flies from north to south in this panel, note the higher tropopause height south of the fold.

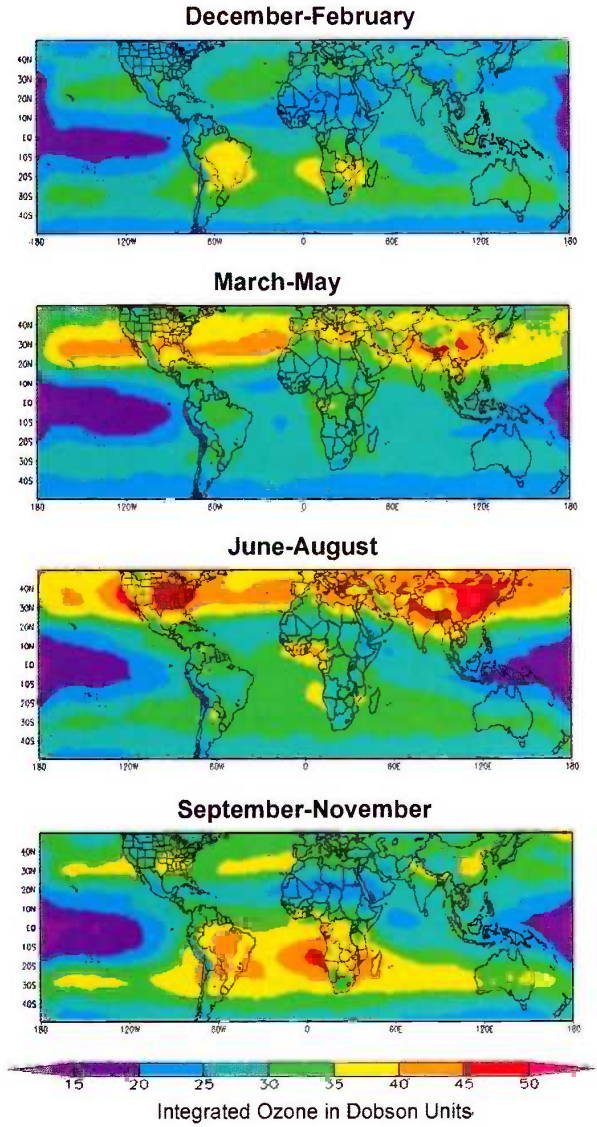


Figure 1 (Chapter 3) Climatological distribution of tropospheric ozone derived from satellite measurements between 1979 and 2000 (from Fishman et al., 2002). Units of contours and Dobson Units (DU). Regions greater than 40 DU have been shaded.

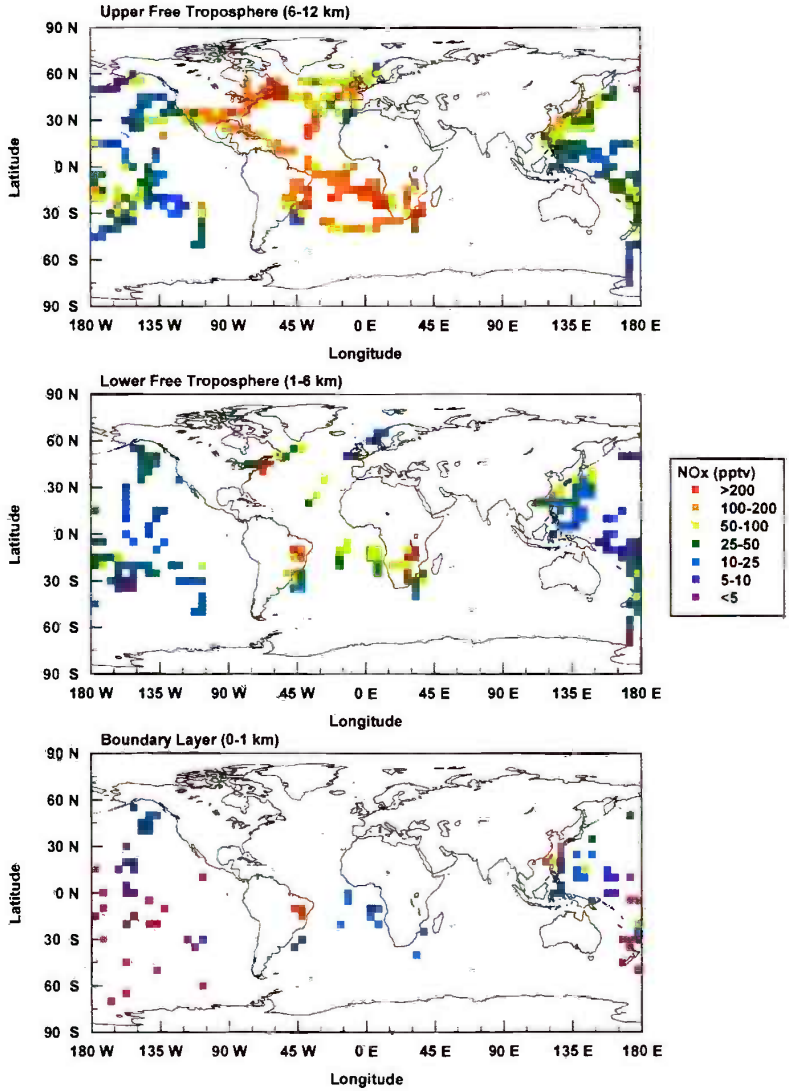


Figure 3 (Chapter 4) Distribution of NO_x based on measurements taken from NASA's DC-8 aircraft during fall (see text for details). Data are averaged on a 5° × 5° latitude–longitude grid for three altitude ranges.

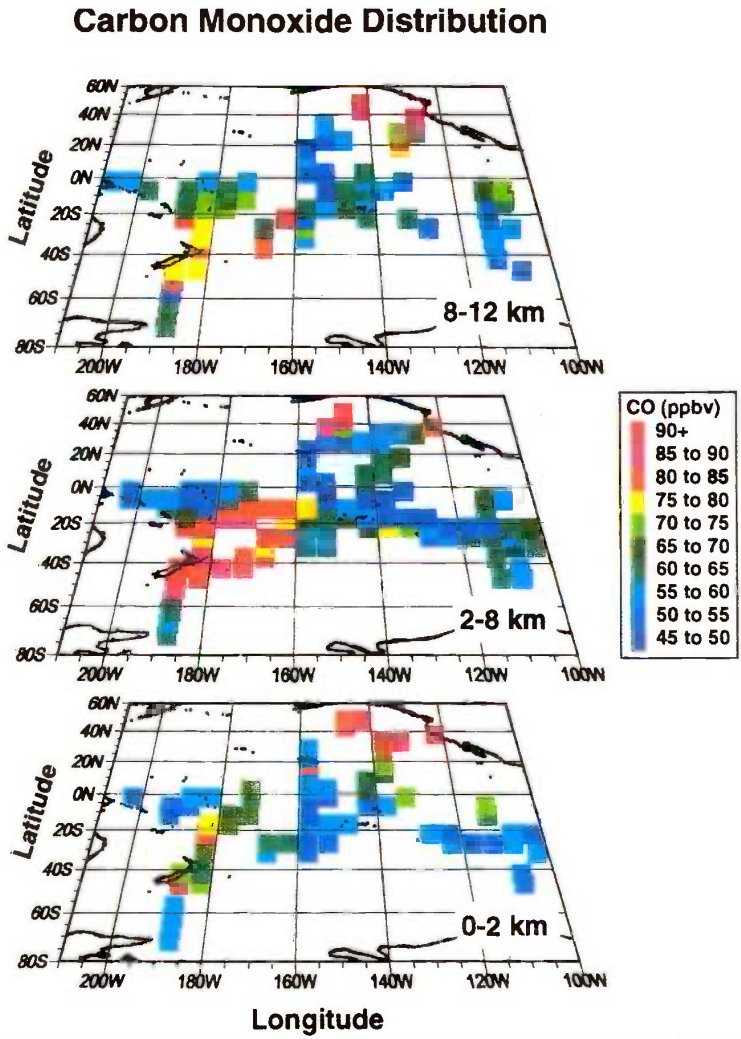


Figure 4 (Chapter 14) CO over tropical Pacific during September 1996 PEM–Tropics A sampling (from Blake et al., 1999). Measurements by G. W. Sachse with a lidar-based instrument. Analysis of possible fire sources is described by Olson et al. (1999).

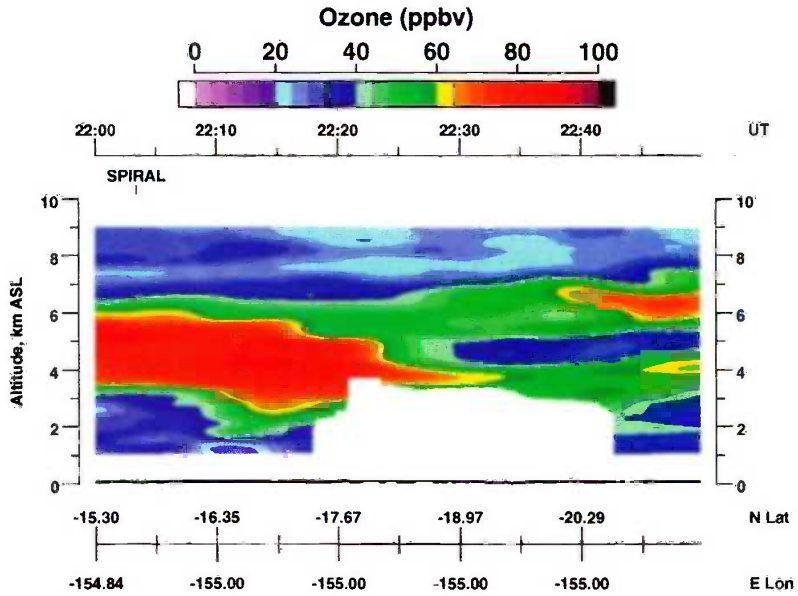


Figure 5 (Chapter 14) Ozone plume over the Pacific seen during the PEM–Tropics A aircraft mission in Sept.–Oct. 1996. (from Fenn et al., 1999).

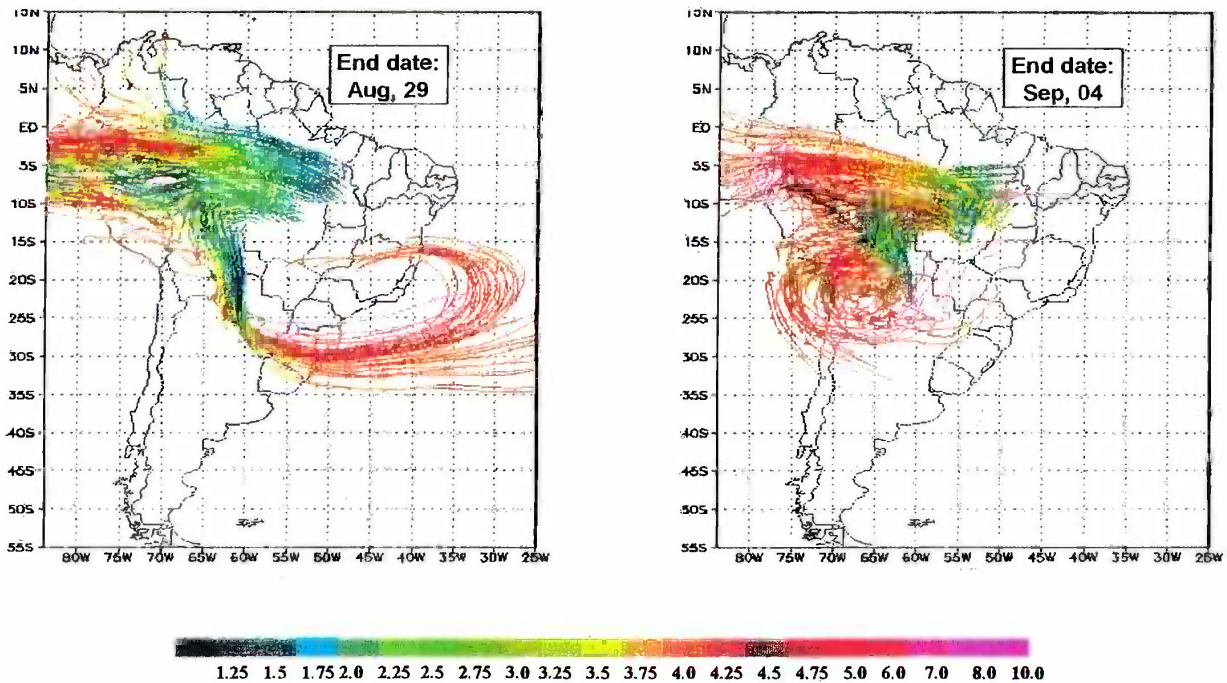


Figure 8 (Chapter 14) Composite of forward trajectories from Cuiabá during the 1995 SCAR-B field experiment. A Brazilian version of the Colorado State mesoscale RAMS model was used to provide winds for the University of São Paulo kinematic trajectory model (from Longo et al., 1999).

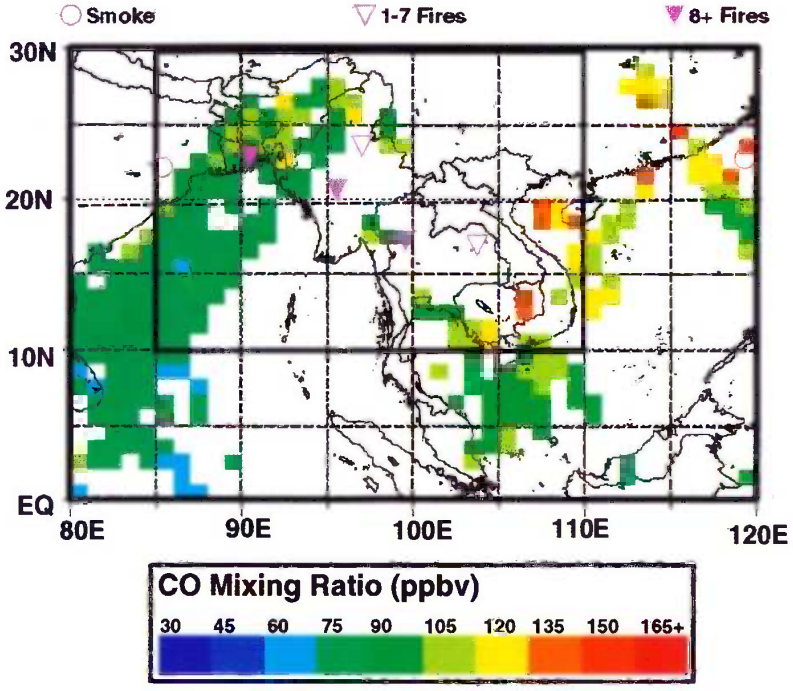


Figure 10 (Chapter 14) (*a*) MAPS CO, April 1994 (from Christopher et al., 1998).

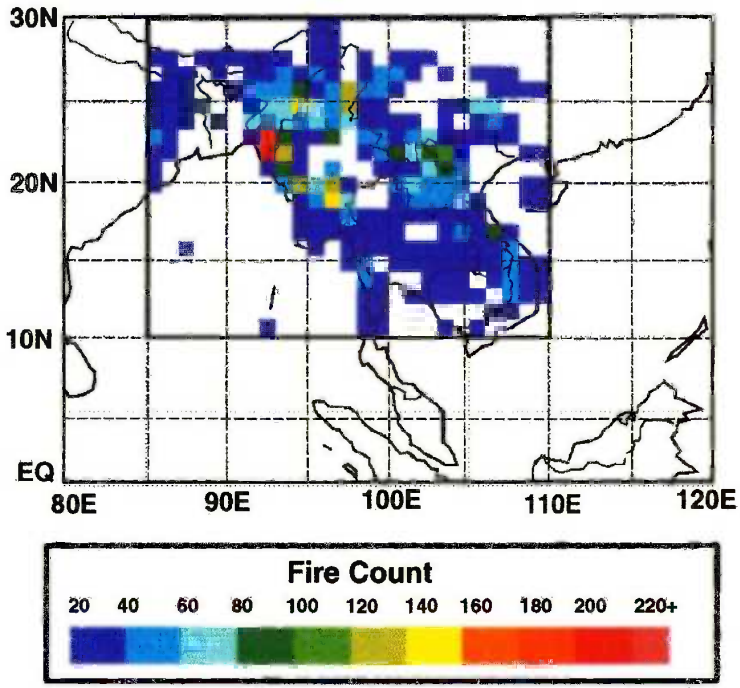


Figure 10 (Chapter 14) (*b*) coincident fires during April 1994 Space Shuttle flight (from Christopher et al., 1998).

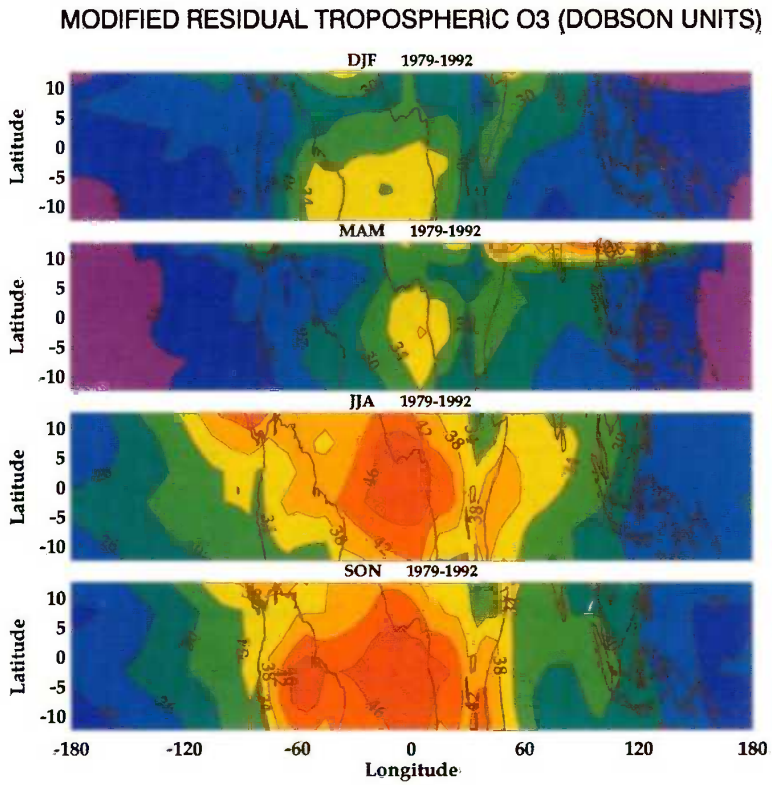


Figure 12 (Chapter 14) Wave-one pattern in tropospheric ozone apparent in TOMS satellite data, averaged from 2 maps/month during the 1979–1992 *Nimbus 7* observing period. Wave appears to be present throughout year. Scale is DU (Dobson units). Cf. Figure A1 in Thompson and Hudson (1999).

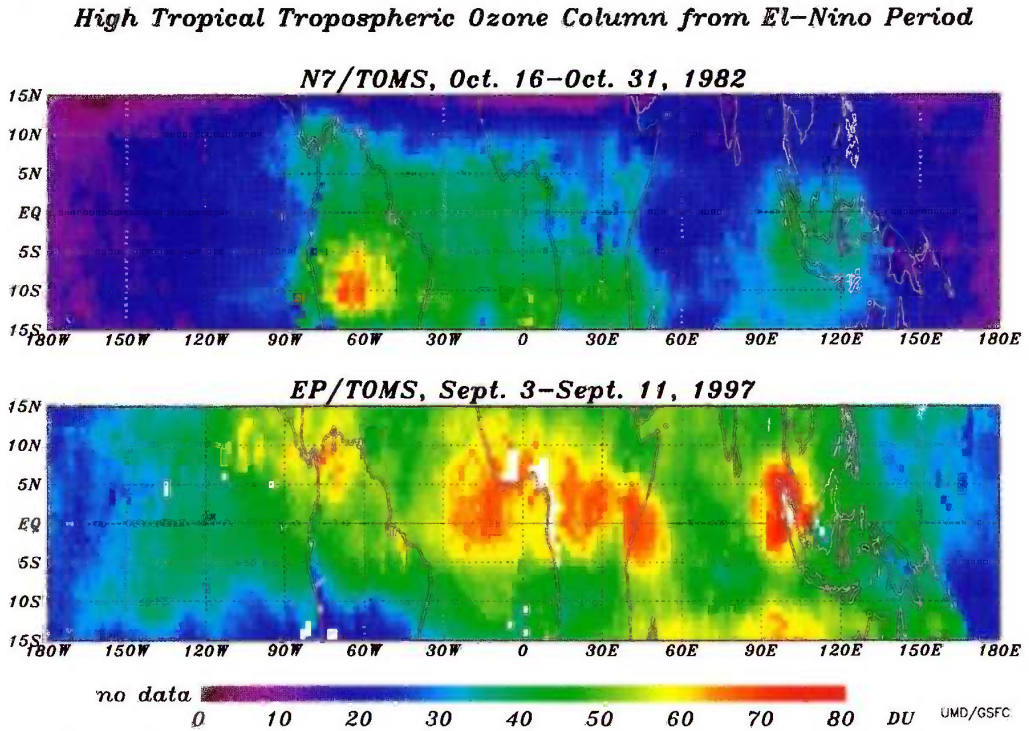


Figure 14 (Chapter 14) Tropospheric column ozone (in DU, from modified-residual method; Thompson and Hudson, 1999) during El Niño-Southern Oscillation (ENSO) of late 1982 (upper panel) as seen in tropical tropospheric ozone map and for September 1997 (lower panel).

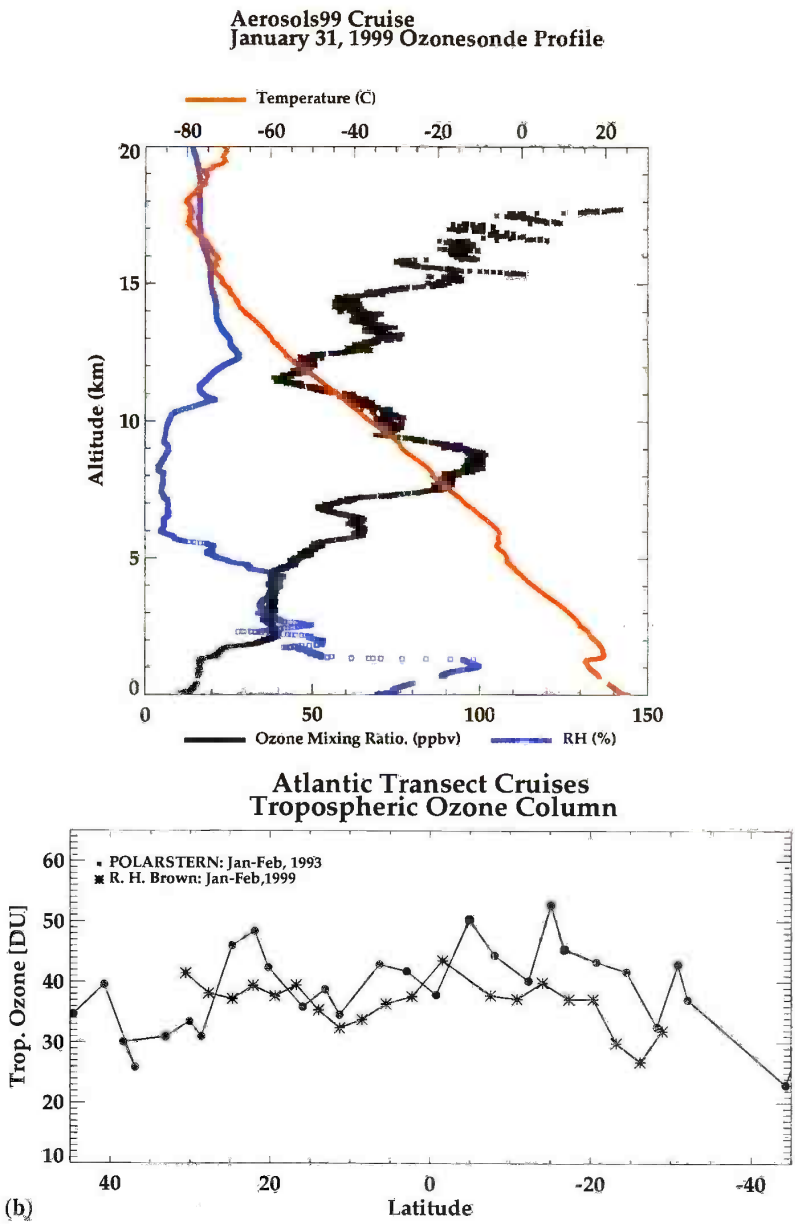


Figure 15 (Chapter 14) (*a*) Profiles of ozone, temperature and water vapor (as percent relative humidity) from 0 to 20 km on January 31, 1999 during Aerosols99 cruise of R/V *Ronald H. Brown*. Anti-correlation of high ozone between 7 and 10 km suggestive of aged stratospheric air. (*b*) Comparison of integrated tropospheric column ozone from sondes launched along Atlantic transect of R/V *Ronald H. Brown* (Thompson et al., 2000) in January–February 1999 and from sondes launched along January–February 1993 Atlantic transect of R/V *Polarstern* (Weller et al., 1996).

Figure 26 (Chapter 16) The brown discoloration resulting from an atmosphere containing nitrogen dioxide (NO_2) being shaded by clouds but viewed against a clear blue sky. Light scattered by particulate matter in the atmosphere can cominate light absorbed by NO_2, causing a gray or blue appearing haze (left side of photograph).

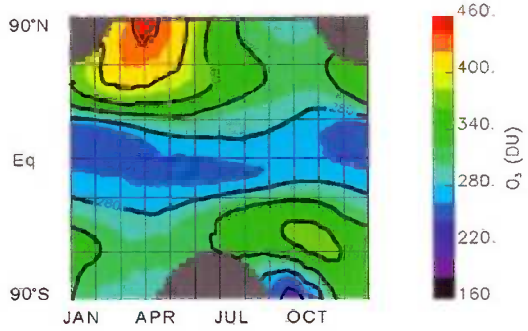


Figure 3 (Chapter 21) Two-dimensional (latitude/season) representation of total column ozone as measured by TOMS for the period 1978 to 1993.

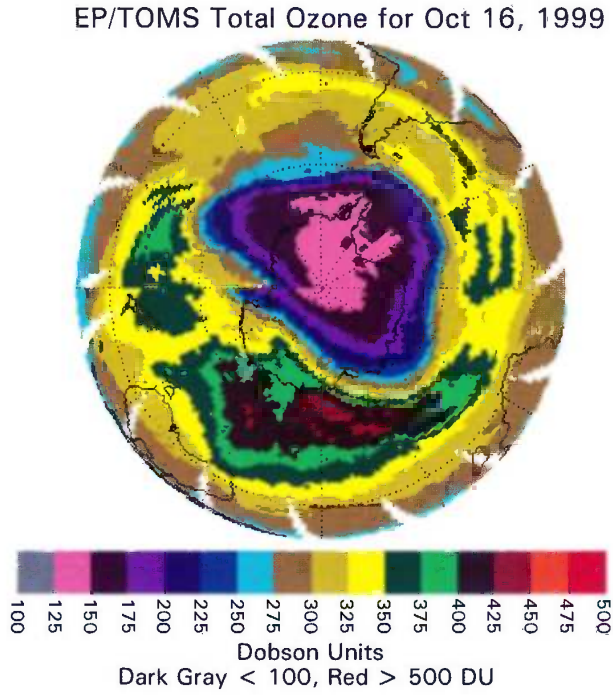


Figure 5 (Chapter 21) Map of total column ozone over the Antarctic as determined from TOMS October 16, 1999.

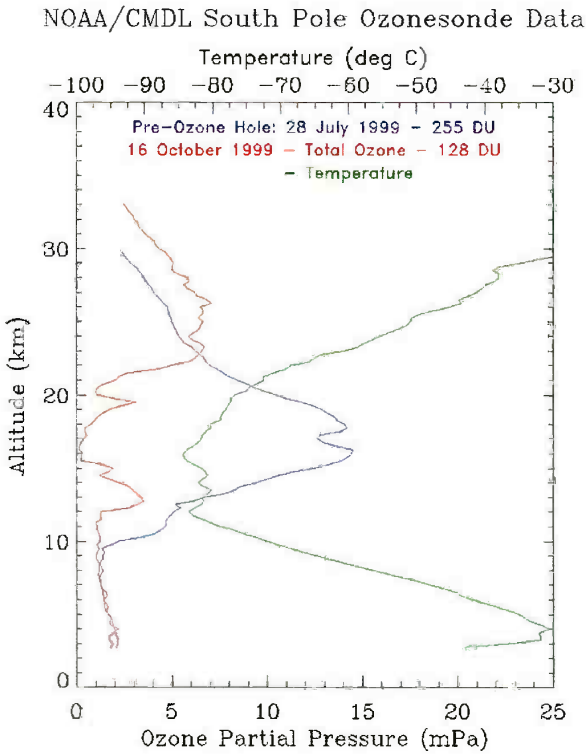


Figure 6 (Chapter 21) Plot of vertical profile of ozone (blue and red lines) over the South Pole as measured from ozonesondes during austral winter (July 28) and spring (October 16), 1999; temperature profile for October 16 is also shown (green line).

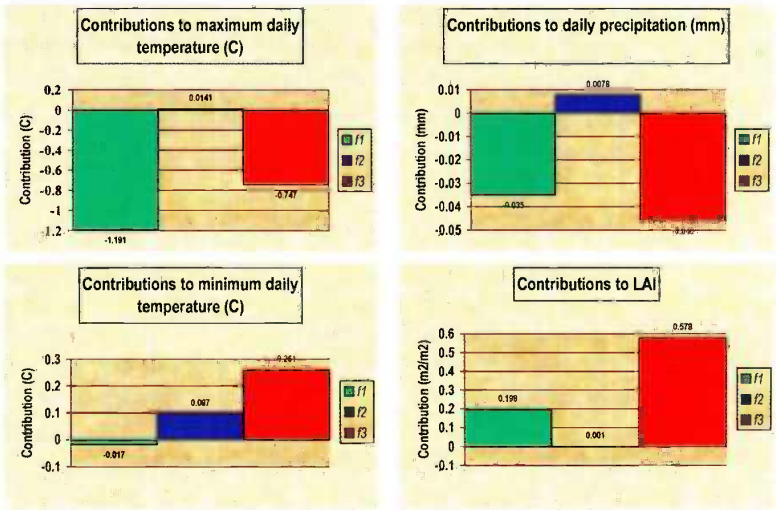


Figure 2 (Chapter 32) RAMS/GEMTM coupled model results—the seasonal domain-averaged (central Great Plains) for 210 days during the growing season, contributions to maximum daily temperature, minimum daily temperature, precipitation, and leaf area index due to f1 = natural vegetation, f2 = 2XC02 radiation, and f3 = 2xC02 biology. (*Adapted from Eastman et al., 2001*).

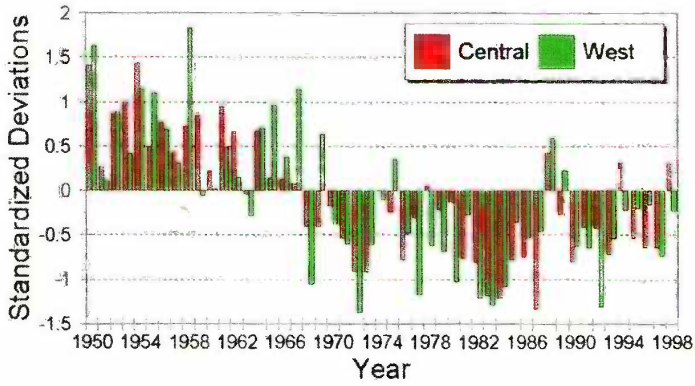
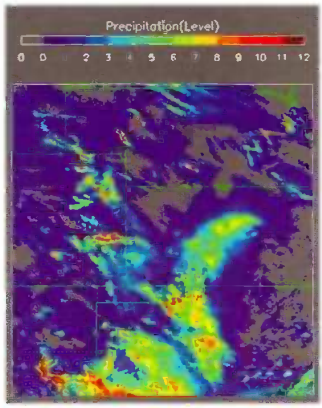
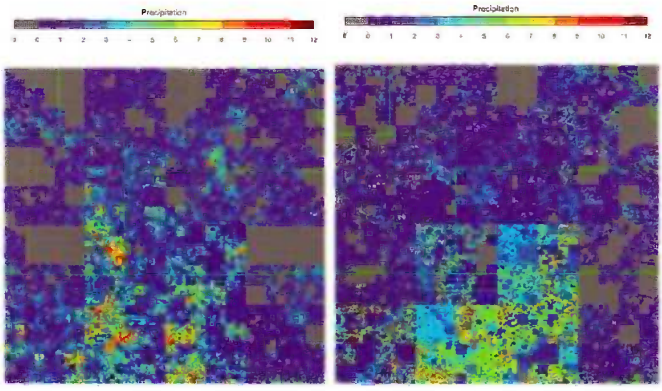


Figure 3 (Chapter 32) Swings or shifts of the time series of standardized deviations of annual rainfall for Central and West Sahel areas during the period 1950–1998. (*After Landsea et al., 1999.*)



Observation



β-lognormal model **Nonparametric hierarchical model**

Figure 6 (Chapter 33) Comparison of observed and downscaled rainfall fields (July 6, 1997) (from Kang and Ramirez[46]).

Figure 1 (Chapter 36) Quebrada San Julián upstream of Caraballeda showing evidence of recent debris flows and flash floods. Note the high slope angles, large numbers of debris flow scars, and abundance of new alluvium and colluvium in the channel bed and fan surface.

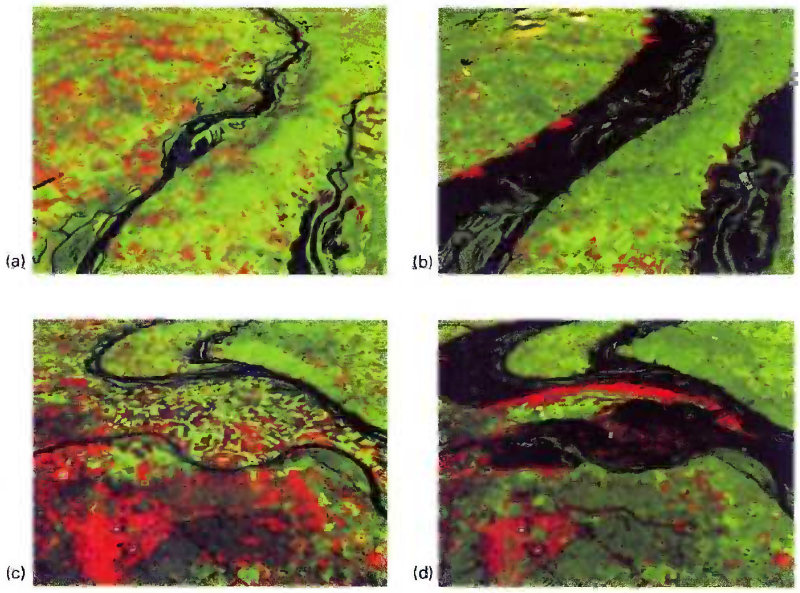


Figure 2 (Chapter 36) (4 panels) These scenes show various sections of the Mississippi River near St. Louis before and just after the 1993 floods, which peaked in late July/early August. The images show the area as seen by the LandSat Thematic Mapper (TM) instrument. The short-wave infrared (TM band 5), infrared (TM band 4), and visible green (TM band 2) channels are displayed in the images as red, green, and blue, respectively. In this combination, barren and/or recently cultivated land appears red to pink, vegetation appears green, water is dark blue, and artificial structures of concrete and asphalt appear dark gray or black. Reddish areas in the scenes during the flood show where water had started to recede, leaving barren land.

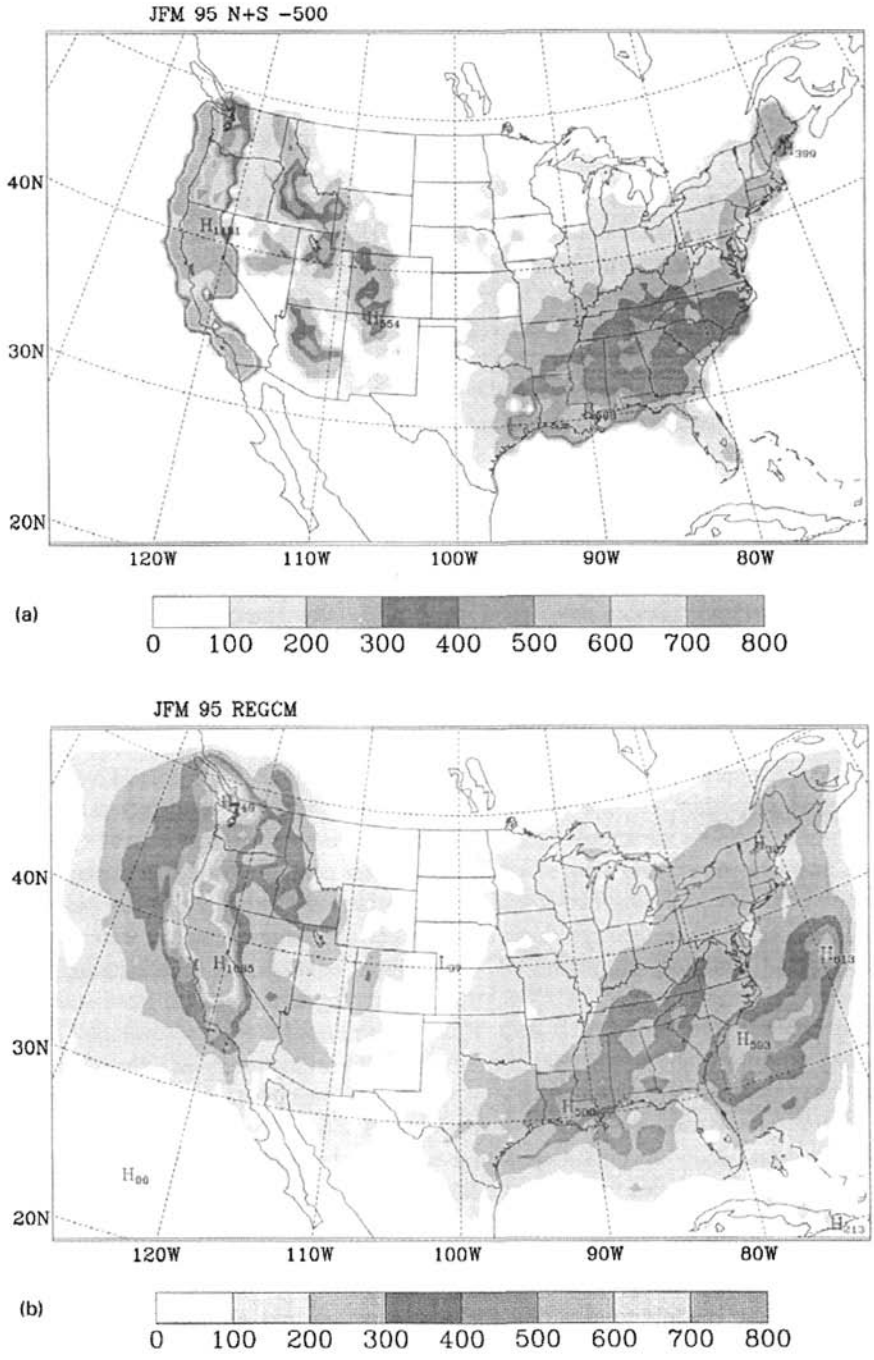


Figure 5 January, February, March 1995 total precipitation (mm) from (*a*) NCDC+SNOW-SNOWTEL; and (*b*) simulated by RegCM.

to AVHRR), but at spatial resolutions ranging from 250 m to 1 km. The sensor has two channels in the visible and near-infrared spectral bands at 250-m resolution, five channels in the visible, near-infrared, and short-wave infrared at 500-m resolution, and the remaining 29 MODIS channels have a spatial resolution of 1 km. The MODIS sensor has on-board visible/near-infrared calibrators, while the AVHRR does not; thus one will be able to derive radiances over snow using some of the MODIS sensors. At least one of the visible MODIS sensors will not saturate over snow. This will be an advancement over the AVHRR and TM sensors that experience significant detector saturation over snow and ice targets in the visible channels. The standard MODIS snow and ice products will consist of 500-m or 1-km resolution binary maps of snow and ice cover, respectively, produced on a global, daily basis in most months.

An approach to modeling spatially distributed snowmelt in steep, alpine basins was proposed using net potential radiation, distributed across the basin using a digital elevation model, as the main factor determining relative snowmelt (Elder et al., 1991b). However, to date this approach has only been applied to small, headwater catchments. It is also possible to infer SWE after the fact from measurements of snow-cover depletion. With a time series of snow cover, e.g., from TM, AVHRR, or MODIS imagery, one can tell when the snow cover disappears, i.e., when snow water equivalence goes to zero. Then using a spatially distributed snowmelt model, one can back calculate from the time snow cover disappears at a point, and then infer the starting value of snow water equivalence. This method has been implemented using TM scenes for a small watershed in the Sierra Nevada, California (Cline et al., 1998).

Research shows that remote sensing also allows estimation of several hydrologic variables important for snowmelt modeling. From airborne hyperspectral sensors [currently NASA's advanced visible/infrared imaging spectrometer (AVIRIS)] one can estimate snow grain size, albedo, liquid water content in the surface layer, and subpixel coverage. Using two-frequency, co-polarized synthetic aperture radar, one can map both snow through thick cloud cover and estimate liquid water content accurately. Work on estimation of snow water equivalence is continuing, with promising results from SIR-C/X-SAR, from photogrammetry and from snowmelt modeling with time-series SCA data. These capabilities have been developed largely using experimental sensors that are not included in currently scheduled satellite launches. A fully automated method of subpixel snow cover mapping uses *Landsat* TM data to map snow cover in the Sierra Nevada and make quantitative estimates of the fractional snow-covered area within each pixel (Rosenthal and Dozier, 1996). Snow fraction estimates from the satellite data can be as accurate as those attainable with high-resolution aerial photography, but they are obtained faster, at much lower cost, and over a vastly larger area.

An important emerging technology is the use of spatially distributed, energy balance snow models to describe the accumulation and ablation of snowpacks. These models organize a wide variety of hydrometeorological and terrain information and permit an improved understanding of snowpack evolution throughout an area of interest. These models have been implemented primarily in well-instrumen-

ted research basins, where high-quality meteorological measurements are available to drive the models. However, recent applications have extended their use to larger regions, where mesoscale numerical weather prediction (NWP) model analyses are used to drive the models (e.g., Cline and Carroll, 1999). The NWS NOHRSC is currently developing a four-dimensional data assimilation system for snow estimation that will use a spatially distributed, physically based snow energy and mass balance model, mesoscale NWP analyses, and an updating scheme to provide operational SWE estimates for the United States.

Gaps in Measurement and Understanding

The most significant constraint on the development of snow hydrology is the general lack of measurements of SWE. Measurements are very sparse throughout the United States, especially in the eastern United States. Most operational in situ SWE measurements are collected to support empirical models that are used to estimate snowmelt runoff. For this purpose, the location of the measurement does not necessarily have to be representative of the surrounding area—it is the relationship between the SWE at particular sites and runoff that is important. This means that most operational SWE measurements (e.g., snow courses and snow pillows) are not reliable indicators of the distribution of SWE in a given area. Rather they are index sites that have snow cover for much of the season, to better support empirical modeling. Furthermore, there are simply too few measurements available to adequately characterize the spatial variability of SWE. The general lack of SWE measurements imposes a significant constraint on the development of improved remote-sensing procedures and distributed snow models, which requires "ground truth" for validation and parameter estimation.

The spatial variability of SWE and snowmelt processes in models needs to be better understood. This is a key area for future development of snow hydrology. Most physical snow process work has been carried out at the point scale, or at very local scales. However, most hydrologic and atmospheric modeling scales are run at much coarser spatial scales that involve significant variability of important processes. These models must parameterize the subgrid variability in some manner, but little effort has been devoted to this problem in snow hydrology. Improved understanding of the variability of SWE and snowmelt processes is also needed to design improved sampling strategies for field measurements. High spatial resolution remote-sensing observations of SWE and other snowpack characteristics using SAR would significantly improve our understanding of the spatial variability of snow properties.

4 ESTIMATION OF SNOWMELT RUNOFF

Historical Approach for Operational Forecasts

Both conceptual and physical approaches have been employed in snowmelt runoff modeling. Conceptual models propose a mathematical relationship between snow-

melt and measured quantities; thus melt can be calculated without treating in detail all of the physical processes and parameters that affect snowmelt. Conceptual models have the benefit of requiring less informational input but suffer from the uncertainty that the conditions hypothesized under different model scenarios are modeled sufficiently.

Operationally, the NWS is tasked with forecasting streamflow, floods, and seasonal water supplies in the United States. Various operational data are used by the algorithms making up the NWS River Forecast System (NWSRFS) to produce streamflow and water supply projections that extend several hours to months into the future. Snow accumulation and ablation is modeled in NWSRFS using an empirical approach that is driven using inputs of air temperature and precipitation (Anderson, 1973; Day, 1990). During periods of precipitation, a simplified energy balance approach is used to estimate snowpack state conditions. When no precipitation is occurring, a simple temperature index method is used. The models are spatially lumped, that is, they model the "average" snowpack state conditions over entire basins or subbasins.

Short-term streamflow forecasts are made using NWSRFS and meteorological forecasts of temperature and precipitation. Because meteorological forecasts become less reliable the further out they extend, this method of streamflow prediction is limited to periods of a few hours to a few days. Beyond that time, the short-term meteorological forecasts are blended into the climatologically average conditions. The long-range forecasting component of NWSRFS, called the extended streamflow prediction (ESP) technique, uses present-day streamflow, soil moisture, and snowpack conditions along with a historical time series of precipitation and temperature to estimate streamflow weeks or months into the future.

Two major factors contribute to high uncertainty in estimates of snowpack conditions, particularly in mountainous regions. First, the NWSRFS snow models contain several empirical parameters that must be calibrated in conjunction with the rest of the NWSRFS algorithms. The empirical calibrations are useful, as they help overcome certain problems, such as poorly representative temperature or precipitation measurement sites. The snow models perform best when conditions are near the average conditions used in the calibration. During extreme or unusual conditions, the models often produce spurious results. Second, accurate estimation of precipitation is critical for these models, but this is especially challenging in mountainous regions. Like the distribution of SWE, the distribution of precipitation is typically highly variable in mountain regions, and there are too few measurement sites to adequately represent this variability. The problem is compounded by the fact that precipitation gages do not measure snowfall well. Consequently, large uncertainties in precipitation inputs to the snow models propagate to the estimates of snow state.

Since current snow state conditions in the model serve as initial conditions for streamflow forecasts, the large uncertainties in modeled snowpack state conditions must be reduced where possible prior to extending the model forward in time. This is accomplished by updating the SWE in the snow models with observed SWE.

In the western United States, two data assimilation schemes are currently used to provide SWE updates for the snow models. The first is a relatively simple approach

that uses satellite observations of snow cover to indicate areas without snow, plus an interpolation of all available surface- and airborne-based SWE observations. The second data assimilation scheme, SEUS (McManamon et al., 1993), assimilates surface, airborne, and satellite snow information in an optimal interpolation framework. Grid points in a basin are classified and the amount of SWE and snowmelt for each class for each year in a historical record is estimated. The historical mean SWE fields are then used to estimate the actual SWE values from current gridded data.

Spatially Distributed Modeling of Melt and Runoff

While empirical snowmelt runoff models have traditionally been useful for operational runoff volume forecasts, they provide little information on the timing, rate, or magnitude of discharge, and they are inappropriate in situations outside the boundary conditions governing the development of the relevant empirical parameters. Thus they may fail to adequately predict water yield in extreme or unusual years, and they cannot be reliably used in investigations examining snowmelt responses to climate variability and change. These problems, and the increasing importance of understanding intrabasin snowmelt for environmental analysis of such factors as basin ecology (Baron et al., 1993), water chemistry (Wolford et al., 1996), and hillslope erosion (Tarboton et al., 1991) have motivated the development of physically based, spatially distributed snowmelt models in recent years. Such models require information on the spatial distribution of snowpack water storage. But mountain snowpacks are spatially heterogeneous, reflecting the influences of rugged topography on precipitation, wind redistribution of snow, and surface energy fluxes during the accumulation season (Elder et al., 1991a), and no widely suitable method yet exists to directly map SWE or simulate distributed snowmelt in rugged mountain regions. One of the main obstacles to physically based modeling is the compilation of the necessary meteorological and snow-cover data to run, calibrate, and validate such models. For example, basin discharge has frequently been used as the sole physical criterion of model calibration and performance assessment for conceptual snowmelt models. But as it is an integrated response to melt and runoff, basin discharge is not sufficient to discriminate between the effects of the multiplicity of data inputs driving physical models and that distributed snow-cover data are required to assess model performance (Bloschl et al., 1991a, 1991b).

A distributed snowmelt model consists of two general components: a model for calculating melt at a single point (given a set of prescribed snowpack and meteorological conditions), and a method of developing the requisite snowpack and meteorological data for all points within the basin. Snowmelt modeling efforts require several steps necessary to couple basinwide energy balance snowmelt models with remote sensing and flow routing. The model topographic distribution of solar radiation (TOPORAD) (Dozier and Frew, 1990) uses information on watershed topography (i.e., a digital elevation model) to spatially distribute radiation. Simpler models are used to spatially distribute other energy balance components. These radiation maps are then used as inputs to models to estimate the distribution of

snow around the basin prior to snowmelt (i.e., initial conditions for melt), and as inputs to the point snowmelt model.

Climate Change and Variability Issues

Expected changes in global climate are cause for concern for future water resources in the western United States, where limited water supplies are already in great demand, and a constrained water regulatory system struggles to satisfy water users. The main question for the regions is: What does potential climate change mean to snowmelt-dominated western water resources? Certainly the assumption of stationarity of the present snowmelt and streamflow regime must be questioned (Dracup and Kendall, 1990); however, it is not clear what exactly should take its place. One criterion with which many global climate models are run is the anticipated doubling of greenhouse forcing from atmospheric greenhouse gases; this doubling and the resulting modeled climate changes are expected to occur by the mid-twenty-first century, well within the "design life" of many existing water control structures and regulations, and indeed within the careers of the next generation of water planners. For this reason, the problem cannot be easily ignored; however, rapid advances in understanding of the potential effects of climate change on snowmelt-dominated hydrologic systems are necessary before an informed response by water resource planners can be made.

A variety of approaches have been used to estimate the effects of or sensitivity to climate change in snowmelt-dominated basins in the western United States. The methods range from purely statistical techniques (Duell, 1992), regression models, to complex deterministic spatial modeling strategies that explicitly account for every process in as detailed a manner as possible (Leavesley et al., 1992). Potential effects of climate change on snowmelt and streamflow in the west include reduced annual streamflow, earlier peak flows in spring, and less winter precipitation occurring as snow. However, due to the great uncertainty in the input data, and the general inapplicability of the models to forecasts beyond the range of their prior calibration, magnitudes remain largely unknown. There are also issues with the variety of models in use, and few cases where a common approach has been employed in more than one hydroclimatic region.

Two points illustrate the tentative and qualitative nature of work to date and highlight the need for a more rigorous approach. First, examination of the major topographic effects on temperature and precipitation in the western United States shows that only regional-scale climate models are appropriate for use in this part of North America. The results of any modeled climate change scenario for the western United States that has not incorporated reasonable representations of topography must be viewed as spurious at best (Seth et al., 1999). In addition to, and because of, the lack of adequate topographic representation in general circulation models (GCMs), several of the important consequences of topographically controlled precipitation and temperature, such as albedo differences due to the presence of snow at high elevations, and forced uplift and cooling of large-scale airflow over individual mountain ranges are also missing from GCMs. This suggests that it may not be

feasible to treat GCM results as even a general trend or rough estimate of realistic climate changes. This is not so much a criticism of GCMs in general, but rather a criticism of the use of their output for modeling the hydrology of basins with topography of which the GCMs are not even aware. Thus accurate quantitative climate change scenarios are conspicuously absent from exercises attempting to predict climate change effects on streamflow in the west.

The second point of the argument is that hydrologic simulations are severely limited by the excessive need to calibrate snowmelt runoff models to achieve a good fit to a basin hydrograph. Under the assumption of stationarity, such calibration is acceptable if runoff is the only variable of interest. Unfortunately, there is absolutely no reason to assume that if climate changes the hydrologic characteristics of a basin will remain stationary. If internal basin hydrologic characteristics are also of interest, then most existing snowmelt runoff models cannot be used since they do not provide any information on internal basin processes. In fact, most of the models depend on sufficient aggregation of internal processes so that accurate outflow predictions may be made. Distributed parameter models appear to address the internal basin processes to a larger degree, but often require more extensive data.

In most studies to date of the potential effects of climate change on snowmelt runoff, the "effect" receiving the greatest attention is the basin hydrograph. Internal basin characteristics should be given greater attention for two reasons. First, basins themselves are important, and we should be concerned about how they may be affected by changing climate. Second, we need to begin incorporating the notion of nonstationarity of basin properties into hydrologic models. For example, climate-change-induced changes in the extent and characteristics of forest cover within a basin might have a greater effect on both the timing and magnitude of snowmelt runoff than changes in temperature or precipitation. Changes in the proportion of rainfall/snowfall could lead to important changes in sediment transport, mass wasting, and other geomorphic characteristics of a basin, which would affect appropriate selection and use of hydrologic models. The hydrochemistry of streams and rivers is largely dependent on the interaction of water with various basin component; thus intrabasin biological and geomorphic changes due to climate change could have important implications for stream water chemistry. Improvements in distributed-parameter hydrologic models and particularly in methods of collecting sufficient input data for these models will certainly be necessary before internal basin processes can be adequately examined.

REFERENCES

Anderson, E. A., *National Weather Service River Forecast System: Snow Accumulation and Ablation Model*, NOAA Technical Memorandum, NWS, Hydro-17, U.S. Department of Commerce, Washington, DC, 1973.

Baron, J. S., L. E. Band, S. W. Running, and D. Cline, The effects of snow distribution on the hydrologic simulation of a high elevation Rocky Mountain watershed using Regional HydroEcological Simulation Systems, RHESSys, *Eos, Trans. Am. Geophys. Union*, *74*(43), 237, 1993.

Bloschl, G., D. Gutknecht, and R. Kirnbauer, Distributed snowmelt simulations in an Alpine catchment 1. Model evaluation on the basis of snow cover patterns, *Water Resour. Res.*, *27*(12), 3171–3179, 1991a.

Bloschl, G., D. Gutknecht, and R. Kirnbauer, Distributed snowmelt simulations in an Alpine catchment 2. Parameter study and model predictions, *Water Resour. Res.*, *27*(12), 3181–3188, 1991b.

Cline, D., K. Elder, and R. Bales, Scale effects in a distributed snow water equivalence and snowmelt model for mountain basin, *Hydrol. Process.*, *12*(10–11), 1527–1536, 1998.

Cline, D., and T. Carroll, Inference of snow cover beneath obscuring clouds using optical remote sensing and a distributed snow energy and mass balance model, *J. Geophys. Res. Atmos.*, *104*(D16), 19631–19644, 1999.

Day, G. N., *A Methodology for Updating a Conceptual Snow Model with Snow Measurements*, NOAA Technical Report, NWS 43, U.S. Department of Commerce, Washington, DC, 1990.

Dozier, J., and J. E. Frew, Rapid calculation of terrain parameters for radiation modeling from digital elevation data, *IEEE Trans. Geosci. Remote Sensing*, 28, 963–969, 1990.

Dracup, J. A., and D. R. Kendall, Floods and droughts, in P. E. Waggoner (Ed.), *Climate Change and U.S. Water Resources*, Wiley, New York, pp. 243–267, 1990.

Duell, L. F. W., Jr., Use of regression models to estimate effects of climate change on seasonal streamflow in the American and Carson River Basins, California-Nevada, in Hermann, Raymond, ed., Managing water resources during global change: 28th Annual Conference, American Water Resources Association, Reno, Nev., November 1992, Proceedings, p. 731–740.

Elder, K., R. E. Davis, and R. C. Bales, Terrain classification of snow-covered watersheds, *Proc. Eastern Snow Conf.*, 48, 39–49, 1991a.

Elder, K., J. Dozier, and J. Michaelsen, Snow accumulation and distribution in an alpine watershed, *Water Resour. Res.*, 27, 1541–1552, 1991b.

Hart, D., and F. Gehrke, Status of the California Cooperative Snow Survey Program, in *58th Western Snow Conference Proceedings*, Sacramento, CA, 1990, pp. 9–14.

Leavesley, G. H., M. D. Branson, and L. E. Hay, Using coupled atmospheric and hydrologic models to investigate the effects of climate change in mountainous regions, in Hermann, Raymond, ed., Managing water resources during global change: 28th Annual Conference, Americal Water Resources Association, Reno, Nev., November 1992, Proceedings, p. 691–700.

Lin, H. F., F. K. Hare, and K. P. Singh, Influence of the atmosphere, in M. G. Wolman and H. C. Riggs (Eds.), *Surface Water Hydrology: Boulder, Colorado*, Geological Society of America, The Geology of North America, 1990, pp. 11–53.

McManamon, A., T. L. Szliga, R. K. Hartman, G. N. Day, and T. R. Carroll, Gridded snow water equivalent using ground-based and airborne snow data, in *Proceedings of the 50th Eastern Snow Conference*, Quebec City, 1993, pp. 75–81.

Matson, M., C.F. Roeplewski, and M.S. Varnadore, 1986: An Atlas of Satellite-Derived Northern Hemisphere Snow Cover Frequency, National Weather Service, Washington D.C., 75 p.

Matson, M., 1991: NOAA satellite snow cover data, Palegeography and Paleoecology, 90: 213–280

Paulson, R.W., E.B. Chase, R.S. Roberts, and D.W. Moody, compilers, *National Water Summary 1988–89: Hydrologic Events and Floods and Droughts*, U.S. Geological

Survey Water Supply Paper 2375, U.S. Government Printing Office, Washington, DC, 1991, p. 591.

Rosenthal, W., and J. Dozier, Automated mapping of montane snow cover at subpixel resolution from the Landsat Thematic Mapper, *Water Resour. Res.*, *32*(1), 115–130, 1996.

Seth, A., R. C. Bales, and R. E. Dickinson, A framework for the study of seasonal snow hydrology and its interannual variability in the alpine regions of the Southwest, *J. Geophys. Res.*, *104*(D18), 22117–22135, 1999.

Shi, J. and J. Dozier, Inferring snow wetness using C-band data from DIR-C's polarimetric synthetic aperture radar, *IEEE Trans. Geosci. Remote Sensing*, *33*(4), 905–914, 1995.

Shi, J., and J. Dozier, Estimation of snow water equivalence using SISR=C/X-SAR, in *Proceedings IGARRS 96* , IEEE No. 96Ch35875, 1996, pp. 2002–2004.

Shi, J., and J. Dozier, Mapping seasonal snow with SIR-C/X-SAR in mountainous areas, *Remote Sensing Environ.*, *59*, 294–307, 1997.

Shi, J. and J. Dozier, Estimation of Snow Water Equivalence Using SIR-C/X-SAR Image Data, Proceedings Progress in Electromafnetic Research Symposium, p. 1041, 1995.

Tarboton, D. G., M. J. Al-Adhami, and D. S. Bowles, Preliminary comparisons of snowmelt models for erosion prediction, *Proc. Western Snow Conf.*, *59*, 79–90, 1991.

Wolford, R. A., R. C. Bales, and S. Sorooshian, Development of a hydrochemical model for seasonally snow-covered alpine watersheds: Application to Emerald Lake Watershed, Sierra Nevada, California, *Water Resour. Res.*, *32*(4), 1061–1074, 1996.

Xu, H., J. O. Baily, E. C. Barrett, and R. E. J. Kelly, Monitoring snow area and depth with integration of remote-sensing and GIS, *Int. J. Remote Sensing*, *14*(17), 3259–3268, 1993.

CHAPTER 26

EVALUATING THE SPATIAL DISTRIBUTION OF EVAPORATION

WILLIAM P. KUSTAS, M. SUSAN MORAN, AND JOHN M. NORMAN

1 INTRODUCTION

Evaporation of water from soil and plant surfaces forms the connecting link between the energy balance and the water balance at Earth's surface. This phenomenon influences the large-scale circulation of the planetary atmosphere, affects soil moisture content that in turn affects hydrologic response, and regulates the microscale carbon dioxide uptake of stomata in individual plant leaves. The vast range of scales encompassed by the process of evaporation makes it of vital environmental interest.

Over the past century, theoretical, modeling, and experimental efforts have greatly expanded our ability to evaluate water loss due to evaporation at local scales using conventional instrumentation. In recent decades, a concerted effort has been made to develop techniques for evaluating the spatial distribution of evaporation at regional and global scales. This effort has been largely focused on the use of remotely sensed information available from sensors aboard orbiting satellite platforms. The result has been a variety of methods that vary in complexity from statistical approaches to physically based analytical approaches and ultimately to numerical process models that simulate the flow of heat and water through the soil, vegetation, and atmosphere.

This chapter will present a brief discussion of the physics of evaporation, highlight conventional methods for estimating evaporation rates, and then will focus on the use of remote sensing for evaluation of the spatial distribution of evaporation at the local, regional, and global scales. Emphasis will be placed on methods for estimating evaporation at an hourly to daily time frame, which is most appropriate for atmospheric, hydrological, and agricultural applications. This work will conclude

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

with a synthesis of the most important research and development issues related to the implementation of such approaches on an operational basis. Although much of the material in Sections 4 and 5 is from the work of Kustas and Norman (1996), new information and results from more recent studies are included.

2 SHORT HISTORY

Although the evaporation process has intrigued humankind for centuries, progress in understanding the physics of evaporation remained slow until the twentieth century when Bowen (1926) showed how the partitioning of available energy between the fluxes of sensible and latent heat could be determined from gradients of temperature and humidity:

$$\lambda E = -(R_n + G)/(1 + \beta) \tag{1}$$

where λE* is the latent heat flux (W/m^2), R_n is the net radiation flux at the surface (W/m^2), G is the sensible heat flux conducted to the soil (W/m^2), and β is the Bowen ratio (Table 1). The ratio of sensible heat (H) to latent heat flux density is

$$\beta = H/\lambda E \tag{2}$$

In Eq. (1), fluxes away from the surface are negative and those toward the surface are positive. The Bowen ratio can be derived from temperature and humidity measurements:

$$\beta = \gamma(K_h/K_v)(\Delta T/\Delta e) \tag{3}$$

where γ is referred to as the psychrometric constant (2.453 MJ/kg at 20°C), K_h and K_v are the eddy transfer coefficients for sensible and latent heat, respectively, and ΔT and Δe are the differences in temperature in degrees centigrade and vapor pressure in kilopascals over the same elevation difference, Δz.

Following the work of Bowen (1926), Penman (1948) combined the thermal energy balance with certain aerodynamic aspects of evaporation and developed an

* Evaporation (E) is often represented in units of mm/day or mm/h but can also be expressed in energy units, where E is the evaporation rate ($kg/s\,m^2$), λ is the heat of evaporation (J/kg), and λE is the latent heat flux density (W/m^2). Though expressed in different units, the terms E and λE are interchangeable. To avoid confusion herein, the term E* will represent evaporation rate in units of depth (mm/h or mm/d), E will represent mass flux density ($kg/s\,m^2$ or $kg/d\,m^2$), and λE will represent latent heat flux density (in units of W/m^2 or $MJ^{-2}\,d^{-1}$). For further clarification on evaluation of Eqs. (1) to (9), readers are encouraged to review Table 1, and consult the treatise by Monteith (1981) and the books by Brutsaert (1982) and Jensen et al. (1989).

TABLE 1 Summary of Scientific and Technical Notation

$\acute{\alpha}$	Surface shortwave albedo
α	Priestley–Taylor coefficient, $\alpha = 1.26$ for regions with no or low advective conditions
β	Bowen ratio, where $\beta = H/\lambda E$
C_p	Specific heat at constant pressure (kJ/kg°C)
d_0	Displacement height (m)
γ	Psychrometric constant (in units of MJ/kg or kPa/°C)
γ^*	$\gamma(1 + r_c/r_a)$ (kPa/°C)
ΔT	Difference in temperature (°C) over the elevation Δz
Δe	Difference in vapor pressure (kPa) over the elevation Δz
Δz	Elevation difference (m)
Δ	Slope of the saturation vapor pressure–temperature curve (kPa/°C)
e_z^o	Saturation vapor pressure at the z level above the surface (kPa)
e_z	Actual vapor pressure at the z level above the surface (kPa)
$e_z^o - e_z$	Vapor pressure deficit (kPa)
e'	Instantaneous deviation of the partial water vapor pressure from the mean at height z
E	Mass flux density (kg/s m^2 or kg/d m^2)
E^*	Evaporation rate in units of depth (mm/h or mm/d)
EF	Evaporative fraction, where EF$= -\lambda E/(R_n + G)$
ε_s	Surface emissivity
f_g	Fraction of green or actively transpiring vegetation
f_{gr}	Fraction of green vegetation viewed by the radiometer
G	Soil heat flux density (W/m^2)
H	Sensible heat flux density to the air (W/m^2)
$H + \lambda E$	Turbulent fluxes (W/m^2)
H_c	Sensible heat flux density from the canopy (W/m^2)
H_s	Sensible heat flux density from the soil (W/m^2)
k	von Karman's constant ($\approx$0.4)
K_h, K_v	Eddy transfer coefficients for sensible and latent heat, respectively
λE	Latent heat flux density (W/m^2 or MJ^{-2} d^{-1})
λE_c	Latent heat flux density from the canopy (W/m^2)
λE_p	Potential latent heat flux density (W/m^2)
N	Day length (h)
ρ	Air density (kg/m^3)
$\rho_{\Delta\lambda}$	Surface reflectance factor for the spectral range $\Delta\lambda$
ρ_{NIR}, ρ_{Red}	Surface reflectance factors in the near-infrared (NIR) and red spectrum, respectively
P	Atmospheric pressure (kPa)
r_a	Aerodynamic resistance (s/m)
r_c	Canopy resistance to vapor transport (s/m)
r_s	Resistance to heat flow in the boundary layer immediately above the soil surface (s/m)
R_n	Net radiant flux density at the surface (W/m)
$R_n + G$	Available energy (W/m^2)
R_{nc}	Absorbed net radiant flux density by the plant canopy (W/m^2)

(*continued*)

TABLE 1 (*continued*)

R_s	Incoming shortwave solar radiant flux density (W/m^2)
R_{ld}	Incoming longwave radiant flux density (W/m^2)
R_{lu}	Upwelling longwave radiant flux density, represented by $\varepsilon_s \sigma T_{sh}^4$
σ	Stefan–Boltzman constant (5.67×10^{-8} W/m^2 K^4)
t	Time starting at sunrise (h)
T_a	Air temperature (°C)
T_{aero}	Surface aerodynamic temperature (°C)
T_c	Canopy temperature (°C)
T_{rad}	Radiometric temperature measured by an infrared radiometer from a space-borne platform
T_s	Soil surface temperature (°C)
T_{sh}	Hemispherical radiometric temperature (C or K)
u	Horizontal wind speed (m/s)
u_s	Horizontal wind speed (m/s) about 5 cm above the soil surface
w	Mean vertical wind at height z (m/s)
w'	Instantaneous deviation of vertical wind speed from w (m/s)
W_f	Wind function [generally, $a + b(u)$, where u is the wind speed in m/s]
Φ_h, Φ_m	Stability corrections for heat and momentum, respectively
z	Height above the surface at which u is measured (m)
z_{om}, z_{oh}	Roughness lengths for momentum and heat (m), respectively
subscript i	Instantaneous values
subscript d	Daily values
subscript m	Midday values

equation for estimating evaporation that was soon adopted by hydrologists and irrigation specialists. The general form of the Penman combination equation is

$$\lambda E = -[(\Delta/(\Delta + \gamma))(R_n + G) + (\gamma/(\Delta + \gamma))6.43 W_f(e_z^o - e_z)] \tag{4}$$

where Δ is the slope of the saturation vapor pressure–temperature curve (kPa/°C), γ is the psychrometic constant (kPa/°C), W_f is a wind function [generally, $a + b(u)$, where u is the wind speed in m/s)], e_z^o and e_z are the saturation and actual vapor pressures at the z level above the surface (kPa), and $(e_z^o - e_z)$ is vapor pressure deficit (kPa).

The Penman formula was recast in terms of an aerodynamic resistance and a surface resistance for application to single leaves (Penman, 1953) and vegetation canopies (Rijtema, 1965; Monteith, 1965). This result, now referred to as the Penman–Monteith equation, is probably the most universally used equation for calculating evaporation:

$$\lambda E = -[\Delta(R_n + G) + \rho C_p(e_z^o - e_z)/r_a]/[\Delta + \gamma^*] \tag{5}$$

where ρ is air density (kg/m^3), C_p is specific heat at constant pressure (kJ/kg°C), and the aerodynamic resistance, r_a (s/m) is

$$r_a = \{[\ln((z - d_0)/z_{0m}) + \ln(z_{0m}/z_{0h}) - \Phi_h][\ln((z - d_0)/z_{0m}) - \Phi_m]\}/k^2 u \quad (6)$$

and z is the height above the surface at which u is measured (m), d_0 is the displacement height (m), z_{0m} and z_{0h} are the roughness lengths for momentum and heat (m), respectively, Φ_h and Φ_m are the stability corrections for heat and momentum, respectively, and k is von Karman's constant ($\approx$0.4). The integral stability functions were summarized by Beljaars and Holtslag (1991) for the stable and unstable conditions. The value of γ^* (kPa/°C) in Eq. (5) is a function of r_a and the canopy resistance to vapor transport [r_c (s/m)], where

$$\gamma^* = \gamma(1 + r_c/r_a) \quad (7)$$

Priestley and Taylor (1972) proposed a simplified version of the Penman combination equation for computation of potential evaporation heat flux density (λE_p) for a surface that has minimal resistance to evaporation. Under these conditions, the aerodynamic component was ignored and the energy component was multiplied by a coefficient,

$$\lambda E_p = -\alpha(\Delta/(\Delta + \gamma))(R_n + G) \quad (8)$$

where $\alpha = 1.26$ for regions with no or low advective conditions.

Regional-scale estimates of evaporation have been made using properties of the atmospheric boundary layer (ABL). One approach applies similarity theory to humidity, temperature, and wind in the ABL (Brutsaert and Mawdsley, 1976). Another approach involves the development of simplified conservation equations for the ABL (McNaughton and Spriggs, 1986). This links the surface fluxes to temporal changes in temperature and humidity in the mixed layer. There are problems in employing either approach. The former has difficulties related to the specification of appropriate roughness parameters, especially in heterogeneous terrain, while the latter must develop parameterizations for advection and entrainment processes that commonly exist in the ABL.

3 CONVENTIONAL APPROACHES FOR MEASURING EVAPORATION

Theoretical developments such as those described in the previous section are generally dependent upon experimental data for verification. There are a variety of conventional approaches for measuring evaporation, ranging from simple to complex and having a range of accuracies and spatial scales.

Most simply, evaporation can be measured under field conditions by monitoring the change in soil water storage over a period of time. Though this can be accomplished fairly easily with a neutron soil water probe, this method does not account

for the drainage from the zone sampled or the upward movement of water from a saturated zone into the zone sampled. Discussions of the problems encountered in determining evaporation by soil sampling were presented by Robins et al. (1954) and Jensen and Wright (1978).

Weighing lysimeters are open-top tanks filled with soil in which crops are grown under natural conditions. Evaporation from the contained soil and plants is generally determined either by weighing the entire unit with a mechanical scale or with a counterbalanced scale and load cell; the reduction in the unit's weight over time equals the rate of water transfer to the atmosphere by evaporation. For accurate results, the soil conditions within the lysimeter should be the same as those without, and the lysimeter must be surrounded by the same vegetation that is growing in the lysimeter for a desired radius of about 100 m. A detailed summary of the use of lysimeters for estimation of evaporation can be found in publications by van Bavel and Myers (1962) and Howell et al. (1985).

Commercial instrumentation is available for determining evaporation using an energy balance approach (Bowen ratio) and a mass transfer method (eddy correlation). The Bowen ratio method [based on Eqs. (1) to (3)] allows values of evaporation to be obtained hourly during daylight hours. The accuracy of the method decreases with decreasing flux of water vapor, or when there is low evaporative demand (e.g., at night). A description of the Bowen ratio equipment was provided by Spittlehouse and Black (1980) and Gay and Greenberg (1985).

The eddy correlation method was proposed by Swinback (1951) based on the theoretical description of the mean vertical flux of water vapor:

$$E = (0.622/P)\rho w'e' \tag{9}$$

where P is atmospheric pressure (kPa), w' is the instantaneous deviation of vertical wind speed from the mean vertical wind (w) at height z, and e' is the instantaneous deviation of the partial water vapor pressure from the mean at height z. Evaluation of Eq. (9) is accomplished using vertical anemometers and vapor pressure sensors with short sampling intervals (hundredths of seconds) to determine w' and e' in short, successive periods of time (tenths of seconds). This method is amenable to field use in routine measurements for extended periods, e.g., months or years (Kanemasu et al., 1979).

Other approaches that have been used to measure evaporation rates include the inflow–outflow method for monitoring evaporation from catchments (Holmes, 1984) and portable gas assimilation chambers (Reicosky, 1981). A limitation of all the techniques described in this section is that they yield essentially point values of evaporation and, therefore, are applicable only to a homogeneous area surrounding the equipment that is exposed to the same environmental factors. An evaluation of the spatial distribution of evaporation over large heterogeneous areas would be prohibitive using these conventional point measurement techniques. There are advantages and disadvantages of these conventional methods and the remote-sensing techniques discussed in the following sections. Conventional methods yield data at one location but operate continuously over time. Techniques that utilize remotely

sensed inputs yield data for each resolution element of the sensor, thus spatially distributed values of evaporation, but at only an instant in time.

4 APPROACHES FOR ESTIMATING EVAPORATION USING REMOTE SENSING

An alternative means of estimating the spatial distribution of evaporation is through the use of remotely sensed images, obtained by either aircraft- or spacecraft-based sensors. Images obtained from existing satellite sensors can cover swaths ranging from 60 to 2050 km (at resolutions ranging from 10 m to 1 km) and include information about surface reflectance, temperature, and general backscatter properties (Table 2).

In this section, recent developments in the evaluation of evaporation using remotely sensed images are discussed, with emphasis on several problems that must be resolved before an operational satellite-based system for monitoring areal evaporation from land surfaces can be realized. These methods have been divided into two basic classes: (a) statistical and analytical approaches that calculate H and λE "directly" from the remote-sensing data and (b) modeling approaches that use remote-sensing data to "define" or serve as boundary conditions in the estimation of λE and H.

Determination of λE Directly from the Remote-Sensing Data

Many approaches for determination of λE directly from remote-sensing data use the surface energy balance equation as the primary boundary condition to be satisfied; that is,

$$R_n + G + H + \lambda E = 0 \tag{10}$$

where $R_n + G$ is often termed the available energy and $H + \lambda E$ are the turbulent fluxes. Evaluation of the available energy is relatively straightforward and will be addressed first, followed by the discussion of more complex evaluation of the turbulent fluxes H and λE.

Approaches for Determining Available Energy

A number of approaches using remote sensing have been developed for estimating the available energy components in Eq. (10). Generally, R_n is evaluated in terms of its four radiation components (Sellers et al., 1990), namely,

$$R_n = (1 - \acute{\alpha})R_s + \varepsilon_s R_{\mathrm{ld}} - \varepsilon_s \sigma T_{\mathrm{sh}}^4 \tag{11}$$

where R_s is the incoming shortwave solar radiation ($\mathrm{W/m^2}$), R_{ld} is the incoming longwave radiation ($\mathrm{W/m^2}$), $\acute{\alpha}$ is the surface shortwave albedo, ε_s is the surface

TABLE 2 Some Current Satellite-Based Sensors

Satellite	Sensor	Spectral Region: Reflective (μm)	Spectral Region: Thermal (μm)	Spectral Region: Microwave (GHz)	Pixel Resolution (PR)	Orbital Characteristics	Repeat Cycle	Time of Data Acquisition	Delivery time from acquisition to user (T_D)
GOES-8	Imager	0.52–0.72 3.8–4.0 6.5–7.0	10.2–11.2 11.5–12.5		1 km (visible) 4 km (all others)	Geostationary	Stationary	Every 30 min	Instantaneous at ground station
METEOSAT	VISSR	0.4–1.1 5.7–7.1	10.5–12.5		Acquired at 1 km Archived at 8 km	Geostationary	Stationary	Every 30 min	Instantaneous at ground station
NOAA-12,14	Advanced Very High-Resolution Radiometer (AVHRR-2)	0.58–0.68 0.725–1.1	3.55–3.93 10.5–11.5 11.5–12.5		1.1 km (local area coverage) 4 km (global area coverage)	Near-polar, sun-synchronous	12 h,every 9.2 days	19.30 (ascending) and 07.30 (descending)	Instantaneous at ground station
Landsat-5	Thematic Mapper (TM)	0.45–0.52 0.52–0.60 0.63–0.69 0.76–0.90 1.55–1.75 2.08–2.35	10.4–12.5		30 m (Vis-IR) 120 m (thermal IR)	Near-polar, sun-synchronous	16 days	Midmorning	72 hours at best, generally 2 weeks to 1 month
Landsat-7	Enhanced Thematic Mapper Plus (ETM+)	0.50–0.90 0.45–0.52 0.52–0.60 0.63–0.69 0.76–0.90 1.55–1.75 2.08–2.35	10.4–12.5		30 m (Vis-IR) 60 m (thermal, IR) 15 m (panchromatic)	Near-polar, sun-synchronous	16 days	Midmorning	48 h
SPOT-1 to *SPOT-3*	High Resolution Visible (HRV)	0.50–0.75 0.50–0.59 0.62–0.66 0.77–0.87			10 m (panchromatic) 20 m (multispectral)	Near-polar, sun-synchronous	26 days, and pointing capability provide shorter cycles	Late morning	48 hours at best, generally 2 weeks to 1 month

ERS-1 to *ERS-2*	Active Microwave (AM-I) Along-Track Scanning Radiometer (ATSR)	1.6 3.7	11 12	5.3 VV (C-band)	1 km (optical) 30 m (3 looks, SAR) 100 m (@ radiometric resolution of 1 dB)	Near-polar, sun-synchronous	3 days	Midmorning and late evening	48 h at best, generally 2 weeks to 1 month
RADARSAT	Synthetic Aperture Radar (SAR)			5.3 HH (C-band)	28 m (4 looks, standard product)	Near-polar, sun-synchronous	24 days	Midmorning and late evening	48 h at best, generally 2 weeks to 1 month
JERS-1	OPtical Sensor (OPS) Visible and Near IR (VNIR) Radiometer Short wavelength InfraRed (SWIR) Radiometer Synthetic Aperture Radar SAR)	0.52–0.60 0.63–0.69 0.76–0.86 1.60–1.71 2.01–2.12 2.13–2.25 2.27–2.40		1.275 HH (L-band)	20 m (OPSVNIR and SWIR) 18 m (3 looks, SAR)	Near-polar, sun-synchronous	44 days	Midmorning and late evening	48 h at best, generally 2 weeks to 1 month
Space Imaging	IKONOS	0.45–0.90 0.45–0.52 0.52–0.60 0.63–0.69 0.76–0.90			1 m (panchromatic) 4 m (multispectral)	Inclination 98.1 °, sun-synchronous	1–3 days	Late morning	24–48 h

(*continued*)

TABLE 2 ***(continued)***

Satellite	Sensor	Spectral Region			Pixel Resolution (PR)	Orbital Characteristics	Repeat Cycle	Time of Data Acquisition	Delivery time from acquisition to user (T_D)
		Reflective (μm)	Thermal (μm)	Microwave (GHz)					
Terra	MOderate Resolution Imaging Spectrometer (MODIS-N)	MODIS 0.66–0.87 (2 bands) 0.47–2.13 (4 bands) 0.42–0.94 (12 bands)	3.8–14.2 (17 bands)		MODIS 0.25 km (Visible, NIR) 0.5 km (Vis, NIR, SWIR) 1 km (Vis, NIR, TIR)	Polar orbiting, sun-synchronous	MODIS 1–2 days	10:30	48 h
	Advanced Space-borne Thermal Emission and Reflectance Radiometer (ASTER)	ASTER 0.52–0.86 (3 bands) 1.60–2.43 (6 bands)	8.3–11.3 (5 bands)		ASTER 15 m (Visible, NIR) 30 m (SWIR) 90 m (Thermal)		ASTER VNIR 5 days SWIR&T 16 days MISR 9 days		
	Multiangle Imaging Spectro Radiometer (MISR)	MISR 0.40–0.88 (4 bands)			MISR 240 m, 1.92 km				

emissivity, σ is the Stefan–Boltzman constant ($5.67 \times 10^{-8}\,\mathrm{W/m^2\,K^4}$), T_{sh} is the hemispherical radiometric temperature (K) as defined by Norman and Becker (1995), so that the quantity $\varepsilon_s \sigma T_{sh}^4$ represents the upwelling longwave radiation flux, R_{lu}. The radiometric temperature measured by an infrared radiometer from a space-borne platform, T_{rad}, is assumed to approximate T_{sh}.

Both R_s and $\acute{\alpha}$ have been estimated from Geosynchronous operational environmental satellites (GOES) using empirical/statistical and physically based models (Pinker et al., 1995). On a daily basis, the estimate of R_s from satellite data has an uncertainty of approximately 10%, but at shorter time scales, for example hourly, the uncertainty may be greater (probably on the order of 20 to 30%, on average), especially for partly cloudy conditions (Pinker et al., 1994). Validating R_s at hourly or shorter time scales under partly cloudy skies is especially difficult due to sampling problems associated with the limited network of ground-based measurements typically available from field experiments (Pinker et al., 1994).

Satellite estimates of the contribution of the net longwave flux at the surface have been developed using sounding data (Darnell et al., 1992). The Tiros Operational Vertical Sounder (TOVS) of the National Oceanic and Atmospheric Administration (NOAA) satellites contains infrared and microwave sensors that can be used for estimating both R_{ld} and T_{rad}. Other approaches have utilized meteorological data collected near ground level with semiempirical relationships for estimating R_{ld}, and then used T_{rad} for calculating the upwelling longwave component (Jackson et al., 1987). Sellers et al. (1990) raise the concern that estimating the four components of R_n could lead to error accumulation, especially in estimating the net longwave flux because both R_{ld} and R_{lu} are large components, so the difference would be small and prone to significant uncertainty. This has led some to estimate surface R_n from the top of the atmosphere (TOA) R_n (Pinker and Tarpley, 1988). While it has been shown that there is little correlation between surface and TOA net longwave flux (Harshvardhan et al., 1990), there is a strong correlation between R_s and R_n at the surface. This has lead to statistical approaches using slowly varying surface properties such as surface albedo and soil moisture with remotely sensed estimates of R_s for estimating R_n (Kustas et al., 1994b). Other techniques use narrow-band reflectance data and T_{rad} from aircraft and satellite-based platforms for estimating the upwelling components $\acute{\alpha}R_s$ and R_{lu} and use meteorological data for estimating the downwelling components R_s and R_{ld} (e.g., Moran et al., 1989; Daughtry et al., 1990). Comparisons with ground-based observations at meteorological time scales (i.e., half-hourly to hourly) indicate that the differences are within the uncertainty in the measurements, namely 5 to 10%.

The soil heat flux (G) can be solved as a function of the thermal conductivity of the soil and the vertical temperature gradient. This temperature gradient cannot be measured remotely, hence numerical models solve for G by having several soil layers (Campbell, 1985). This requires detailed information about soil properties. Models using routine weather data may provide satisfactory predictions of soil heat flux (e.g., Camillo, 1989). An alternative approach takes G/R_n as a constant under daytime conditions that varies as a function of the amount of vegetation cover or leaf area index (LAI), which can be estimated by use of remotely sensed vegetation

indices (VI)* (Choudhury et al., 1994). Several studies have shown that the value of G/R_n typically ranges between 0.4 for bare soil and 0.05 for full vegetation cover (Choudhury et al., 1987). Observations (Clothier et al., 1986; Kustas et al., 1993a) indicate that a linear relationship between VI and G/R_n exists, although analytically it has been shown that the relationship should be nonlinear (Kustas et al., 1993a).

Statistical Approaches for Determination of λE

Statistical methods for estimating λE have mainly been developed to predict daily λE using instantaneous remote-sensing observations and assumptions about the relationship between midday H and λE and $R_n + G$. One of the most widely applied approaches, using a T_{rad} observation near midday, was pioneered by Jackson et al. (1977) whereby they observed that daily differences between λE and R_n could be approximated by this linear expression:

$$R_{n,d} + \lambda E_d = A + B(T_{\text{rad},i} - T_{a,i}) \tag{12}$$

where the subscript i and d represent instantaneous and daily values, respectively, A and B are statistical regression coefficients, and T_a is the air temperature (°C) at about 2 m above the surface. A more general form of this expression was proposed by Seguin and Itier (1983) based on theoretical and experimental observations; namely,

$$R_{n,d} + \lambda E_d = B'(T_{\text{rad},i} - T_{a,i})^n \tag{13}$$

where B' was dependent on surface roughness and the value of n depended on stability ($n = 1$ for stable and 1.5 for unstable conditions). A variant of Eq. (13) was introduced by Nieuwenhuis et al. (1985) where they replaced $T_{a,i}$ and $R_{n,d}$ with a reference canopy temperature ($T_{c,i}$) corresponding to conditions of potential λE ($\lambda E_{d,p}$). The linear form of Eq. (12) has been verified experimentally and theoretically (Carlson and Buffum, 1989; Lagouarde, 1991). Carlson et al. (1995) used a soil vegetation atmospheric transfer (SVAT) model to show that a systematic relationship exists between the B' and n parameters in Eq. (13) and fractional cover, which can be estimated with remotely sensed data. Theoretical and experimental work by Lagouarde and McAneney (1992) resulted in the derivation of an equation for estimating daily sensible heat flux (H_d) using T_{rad} measured around the time of the NOAA–AVHRR (advanced very high resolution radiometer) overpass (1400 local standard time) and maximum T_a. The equation is similar in form to Dalton's evaporation equation (see Brutsaert, 1982) and requires the determination of two empirical parameters relating instantaneous to daytime average values of wind speed and surface–air temperature differences. On a daily basis the above techniques appear to have an uncertainty of ±1 mm/day or 20 to 30%.

* Spectral vegetation indices (VI) are a ratio or linear combination of reflectances in the red and NIR wavebands that is particularly sensitive to vegetation amount (Jackson and Huete, 1991) or the amount of photosynthetically active plant tissue in the plant canopy (Wiegand et al., 1991).

The approaches described above attempt to extrapolate "instantaneous" remote-sensing observations of the derived fluxes to daily totals, which is required for many hydrological and agricultural applications. Interest in daily fluxes led Jackson et al. (1983) to develop a procedure using the assumption that the temporal trend in λE would follow the course of solar radiation during the daylight period. They showed that for a clear day the ratio of daily to midday R_s (R_{sm}) could be approximated by an analytical expression:

$$R_{sd}/R_{sm} = 2N/[\pi \sin(\pi t/N)] \tag{14}$$

where N is the daylength in hours, and t is the time starting at sunrise. Several studies have shown this technique can yield satisfactory estimates of λE using the assumed equivalence $\lambda E_d/\lambda E_m = R_{sd}/R_{sm}$ (Brutsaert and Sugita, 1992).

Experimental observations analyzed by Hall et al. (1992) suggest that the evaporative fraction [EF $= -\lambda E/(R_n + G)$] remains fairly constant over the daytime period. With this assumption, an instantaneous estimate of the fluxes and hence EF from a remote-sensing observation would have the potential to provide daily λE as long as one can estimate the daytime average available energy ($R_n + G$). Several studies have found this technique can give reasonable results with differences in daily E* of less than 1 mm/d (Sugita and Brutsaert, 1991; Brutsaert and Sugita, 1992; Hall et al., 1992; Kustas et al., 1994a). The estimates of daily λE derived from either Eq. (14) or from assuming EF is constant, however, should be adjusted for the contribution of nighttime λE. Nighttime λE can be anywhere from 10 to 30% of the daily total (Owe and van de Griend, 1990). This percentage of the daily total will largely depend upon the climate and season. For temperate climates in the summer, 10 to 20% of the daily total is probably typical (Brutsaert and Sugita, 1992).

Recently, Zhang and Lemeur (1995) examined the underlying assumptions of both Eq. (14) and constant EF using the Penman–Monteith equation, and compared the results to measurements from a mixed agricultural and forested region during HAPEX–MOBILHY (Hydrological Atmospheric Pilot Experiment–Modelisation du Bilan Hydrique; see, e.g., André et al., 1986) under clear skies. They found that EF is fairly constant for short vegetation but may not be for forests. Furthermore, the midday values of EF tended to be smaller than the daytime average and the daytime total available energy is required to use this method. Therefore they felt the approach of Jackson et al. (1983) was more suitable since it required only one instantaneous estimate of λE and Eq. (14) to compute daily λE. However, Eq. (14) will only be suitable for clear-day conditions whereas Sugita and Brutsaert (1991) and Kustas et al. (1994a) found that EF was reasonably constant under a wider variety of conditions.

Analytical Approaches for Determination of H and λE

Price (1980) proposed a model for obtaining daily integrated fluxes directly by integrating Eq. (10) over a 24-h period with some simplifying assumptions. The result is an analytical expression for computing daily λE. It requires as primary input

a 24-h max–min difference in T_{rad} and daily average climate data obtained by routine weather station observations (i.e., wind speed, air temperature, and vapor pressure). This model readily lends itself to the NOAA–AVHRR series of satellites, which provide day–night pairs of radiometric surface temperature. Further refinements to the technique were made by Price (1982) resulting in a prognostic model that appears to give appropriate λE values when compared to local estimates using standard meteorological and pan evaporation data. However, the amplitude of the max–min difference in T_{rad} is affected by more than surface soil moisture when vegetation is present and therefore it is less directly coupled to the relative magnitude of λE (Norman et al., 1995a).

Other methods generally compute λE by evaluating R_n, G and H and solving for λE by residual in Eq. (10). At least one radiometric surface temperature observation is required. Unfortunately, most of the approaches that are described below provide only an instantaneous estimate of the fluxes because these models require T_{rad}, which means that only one estimate of λE can be computed during the daytime except when using T_{rad} observations from satellites such as *GOES* or *METEOSAT*.

With R_n and G estimated by the remote-sensing methods described earlier, sensible heat flux is normally computed using the following expression:

$$H = -\rho C_p(T_{aero} - T_a)/r_a \tag{15}$$

where T_{aero} is the surface aerodynamic temperature (°C) (Norman and Becker, 1995) and T_a is the air temperature (°C) either measured at screen height or the potential temperature in the mixed layer (Brutsaert and Sugita, 1991; Brutsaert et al., 1993). The resistance to heat transfer (r_a) is affected by windspeed, atmospheric stability, and surface roughness (Brutsaert, 1982).

Since T_{aero} cannot be measured by remote sensing, it is usually replaced by T_{rad}. For uniform canopy cover, the difference between T_{aero} and T_{rad} is typically less than 2°C (Choudhury et al., 1986; Huband and Monteith, 1986), but for partial vegetation cover the differences can reach 10°C (Kustas, 1990). This has forced many investigators to adjust r_a via empirical methods related to the scalar roughness for heat (Kustas et al., 1989; Sugita and Brutsaert, 1990; Kohsiek et al., 1993) or to use an additional resistance term (Stewart et al., 1994). However, these adjustments to Eq. (15) are not generally applicable because they have not been related to physical quantities causing differences between momentum and scalar transport (McNaughton and Van den Hurk, 1995). This is supported by Sun and Mahrt (1995) who analyzed T_{rad} observations collected over heterogeneous surfaces and found that existing scalar roughness parameterizations for predicting reliable H fluxes with Eq. (15) were not generally applicable. Efforts have been made to develop dual-source models (Norman et al., 1995b; Lhomme et al., 1994; Chehbouni et al., 1996) to account for differences between T_{aero} and T_{rad}, and thus avoid the need for empirical adjustments to r_a. As a result, dual-source models may have broader application for heterogeneous surfaces (Kustas et al., 1996).

In dual-source modeling approaches, the energy exchange is partitioned between the soil/substrate and the vegetation. An example of a dual-source model was

presented by Norman et al. (1995b), based on the assumption that soil surface and vegetation canopy fluxes can be taken in parallel, where

$$H = H_c + H_s = -\rho C_p\{[(T_c - T_a)/r_a] + [(T_s - T_a)/(r_a + r_s)]\} \tag{16}$$

and H_c and H_s are the sensible heat fluxes from the canopy and soil, respectively, r_s is the resistance to heat flow in the boundary layer immediately above the soil surface, and T_c and T_s are the canopy and soil temperatures, respectively. Though a dual-source approach such as that presented in Eq. (16) has the advantage over single-source approaches [represented by Eq. (15)] of accounting for different sources and sinks of energy fluxes, difficulties arise in specifying the resistances to sensible and latent heat transport from the soil and vegetation. However, relatively simple parameterizations have been proposed. For example, Norman et al. (1995b) proposed that the value of r_s be computed from the equation developed by Sauer et al. (1995)

$$r_s = (a + bu_s)^{-1} \tag{17}$$

where u_s is the wind speed (m/s) about 5 cm above the soil surface, estimated with equations of Goudriaan (1977), and $a \approx 0.004$ m/s and $b \approx 0.012$. Further, they proposed that values of T_c and T_s be derived from T_{rad} using the expression

$$T_{\text{rad}} = [f_{\text{gr}}T_c^4 + (1 - f_{\text{gr}})T_s^4]^{1/4} \tag{18}$$

where f_{gr} is the fraction of green vegetation viewed by the radiometer; and that the absorbed net radiation by the plant canopy, R_{nc}, be partitioned between H_c and λE_c according to the Priestley–Taylor approximation (Priestley and Taylor, 1972), where

$$R_{nc} = -H_c/[1 - 1.3f_g\Delta/(\gamma + \Delta)] \tag{19}$$

where f_g is the fraction of green or actively transpiring vegetation.

A recent study by Zhan et al. (1996) compared several single- and dual-source models for computing H with T_{rad} over different land cover types. They showed that models containing the least empiricism to account for the differences between T_{rad} and T_{aero} gave the best results with differences less than 30%, on average. The dual-source model by Norman et al. (1995b) generally gave the smallest differences with measured H fluxes. The average difference was around 20%, which is considered the level of uncertainty in eddy correlation and Bowen ratio techniques for determining the surface fluxes in heterogeneous terrain (Nie et al., 1992).

Another approach to solve this problem relates to performing detailed simulations using microclimate and radiative transfer models that can predict the relationship between T_{rad} and T_{aero} as a function of surface conditions such as vegetation cover or LAI and surface soil moisture and solar zenith and azimuth angles (Prévot et al., 1994). Some preliminary results from the simulations indicate that LAI is a major

factor in determining the order of magnitude of the scalar roughness needed in Eq. (15) if T_{aero} is replaced by T_{rad}. A similar result using a Lagrangian approach was obtained by McNaughton and Van de Hurk (1995) who represented the difference between momentum and scalar transport using an excess resistance term.

The analytical approaches outlined above require an estimate of T_a. Air temperature is not measured in many regions, and where it is measured it only represents local conditions near the site of the measurement and not at each satellite image pixel. With most current satellite observations of T_{rad} at the 0.10- 1-km pixel scale, significant variations in near surface meteorological conditions may exist depending on surface conditions. Methods using satellite data indicate at least $\pm 3°C$ uncertainty in the estimate of T_a when compared to standard weather station observations (Goward et al., 1994). Zhan et al. (1996) showed that two-source models are generally more sensitive to errors in $T_{rad} - T_a$ than to most other model parameters; thus it is a major advantage for a model not to require a measurement of T_a. Kustas and Norman (1997) revised the Norman et al. (1995b) dual-source model for computing the turbulent fluxes without the need for T_a via the use of T_{rad} observations at two sensor viewing angles, $\sim 0°$ and $\sim 50°$ zenith angles. Such viewing angles from a satellite-based platform have been available from the along track scanning radiometer (ATSR) instrument aboard the *ERS-1* satellite (Prata et al., 1990; Prata, 1993). With the ATSR data, there would be no need to extrapolate T_a from a sparse network of meteorological observations to each satellite pixel, a very unreliable approach. Moreover, the model is essentially unaffected by the typical 1 to 2°C error in estimating T_{rad} from satellites. With these two attributes, the model is well suited for computing regional-scale surface fluxes with an ATSR type of sensor.

Other methods avoid the need for estimating T_a on a pixel-by-pixel basis by relying on air temperature in the ABL, which is much more uniform over a region (Brutsaert and Sugita, 1991; Brutsaert et al., 1993). However, the variability of evaporation is more difficult to quantify. Other approaches attempt to use remotely sensed data in the optical wavebands to define variation in meteorological conditions (Bastiaanssen et al., 1998; Gao et al., 1998). It remains to be seen how universal these relationships are for different climates.

Modeling Approaches That Use Remote-Sensing Data to Define Boundary Conditions

Numerical Models. Several numerical models have been developed over the past decade to simulate surface energy flux exchanges using remote sensing data (usually observations of T_{rad}) for updating the model parameters (Camillo et al., 1983; Carlson et al., 1981; Soer, 1980; Taconet et al., 1986). The advantage of these approaches is that the temporal trend of the fluxes can be simulated and periodically updated with the remote-sensing data. Taconet et al. (1986) show the feasibility of using this approach with AVHRR data and, more recently, included the geostationary satellite data (*METEOSAT*) to increase the stability of the model inversion and atmospheric correction of the satellite observations (Taconet and Vidal-Madjar, 1988).

Unfortunately, these models require many input parameters related to soil and vegetation properties not readily available at regional scales. This has prompted some to simplify numerical models in order that remote sensing could potentially be used to estimate most of them (Bougeault et al., 1991). An extreme example of this is given by Brunet et al. (1991) who use an atmospheric boundary layer (ABL) model to calculate regional-scale energy fluxes with a Penman–Montieth equation for parameterizing the energy transport across the soil–vegetation–atmosphere interface. The surface resistance is the main adjustable parameter and is adjusted in order for the model to match the early afternoon infrared surface temperature observation from the NOAA–AVHRR satellite. Preliminary tests using observations under different moisture and crop conditions and surface temperatures from ground-based stations indicate the model adequately simulates the temporal trace and magnitudes of both the energy fluxes and surface temperature.

Numerical models have several advantages over the statistical and analytical approaches. First, they typically better represent the physics of energy transport in the soil–vegetation–atmosphere system. Second, with initial and boundary conditions, they can simulate the energy fluxes continuously. Yet many numerical models still require continuous weather data such as wind speed, air temperature, and vapor pressure, or in the case of atmospheric models that can simulate the near-surface weather, they require radiation data. In practice, few of these models can be used at regional scales with remote-sensing data because of the large amount of vegetation and soils information required to evaluate necessary parameters. Some success in bridging this gap has been achieved by combining a physically based robust model simulating the energy fluxes with remote-sensing data, which provides necessary information for determining key surface parameters in an operational mode (Sellers et al., 1992; Crosson et al., 1993). Two such approaches that appear to have great potential for estimating λE operationally are discussed below in some detail.

Atmospheric Climate Models. An important conceptual step in improving the procedure for estimating soil moisture and the surface energy balance came with the idea of using the time rate of change of T_{rad} from a geostationary satellite such as *GOES* with an atmospheric boundary layer model (Wetzel et al., 1984). By using time rate of change of T_{rad}, one reduces the need for absolute accuracy in satellite sensing and atmospheric corrections, both major challenges. Diak (1990) improved this approach further with a method for partitioning the available energy $(R_n + G)$ into H and λE by using the rate of rise of T_{rad} from the *GOES* satellite and ABL rise from the 12 Greenwich mean time (GMT) synoptic sounding to the 00 GMT sounding. The model is initialized with the 12 GMT sounding of temperature, humidity, and wind speed. Then the surface Bowen ratio (i.e., the ratio of the turbulent fluxes $H/\lambda E$) and the "effective" surface roughness are varied until the predicted 12-h rise in ABL height and T_{rad} match the observations. This effective surface roughness combines the effects of the surface aerodynamic roughness, viewing angle, and fractional vegetative cover. Estimates of surface albedo and emissivity are required by the model.

Diak and Whipple (1993) further refined the model by including a procedure to account for effects of horizontal and vertical temperature advection and vertical motions above the ABL. Sensitivity of the model to the determination of the surface energy balance and to the effective roughness was performed with a case study using data from the Midwest and Great Plains areas in the continental United States. They also verified their model estimates of the surface energy balance with in situ measurements from the FIFE (First ISLSCP Field Experiment; see Sellers et al., 1988) site for 2 days. The model-derived λE values were within 10% of the measurements, suggesting this technique may provide reliable λE estimates at regional scales. Additional comparisons of 12-h averages of sensible heat flux with FIFE observations support the utility of their model (see Fig. 2 from Diak et al., 1995). They also found that temperature advection usually does not significantly impact the surface energy balance estimates given by the model on a daily basis, although for areas that are routinely affected by advection the biasing could impact longer term averages of λE (i.e., at climate time scales).

In a related approach, Anderson et al. (1997) recently developed and tested a two-source surface energy balance model requiring measurements of the time rate of change of surface temperature and an early morning ABL sounding. With this model, many of the problems associated with the use of radiometric surface temperature were avoided. The model accommodated the first-order dependence of the radiometric surface temperature on view angle, avoided the need for atmospheric corrections and precise emissivity evaluation, and did not require in situ measurements of air temperature. The performance of the model was evaluated with experimental data from FIFE and from a semiarid rangeland experiment (Monsoon'90; see Kustas and Goodrich, 1994). The model yielded uncertainties in flux estimates comparable to models needing in situ air temperature observations and were comparable to the uncertainties in surface energy flux measurements.

Recognizing the fact that using T_{rad} requires detailed information on the characteristics of the surface and the structure of the overlying atmosphere, which is often incomplete for many regions, Diak et al. (1994) have proposed a method that employs the High Resolution Interferometer-Sounder (HIS) for estimating the turbulent heat fluxes, H and λE. The premise is that the temporal changes in the radiances observed by the HIS implicitly measure changes in the lower atmosphere, which are a measure of the absolute amount of energy added to the ABL. The HIS radiance changes were described by coefficients obtained by an eigenvalue decomposition procedure. These coefficients were in turn related to various components of the surface energy balance equation using multiple linear regression. Diak et al. (1994) provide convincing evidence that this method responds to temperature changes in the lower atmosphere as well as surface temperature changes. Consequently, this method is equivalent to the method of Diak (1990), but without requiring any ancillary data, just two remote radiance measurements. However, even when HIS becomes operational, co-located flux measurements will be required to establish a database to use the HIS technique. One possible solution is to identify sites that have sufficiently detailed surface information to permit some of the other techniques described above to be used to calibrate this procedure. In any event, the HIS tech-

nique offers tremendous potential since it can evaluate the surface energy balance relying only on remotely sensed data.

Alternative Approach: Exploiting the VI/T_{rad} Relation. Numerous studies have found a significant negative correlation between the normalized difference vegetation index (NDVI) and T_{rad} over a variety of surfaces (Goward et al., 1985; Hope and McDowell, 1992; Nemani and Running, 1989; Nemani et al., 1993), where

$$\text{NDVI} = (\rho_{\text{NIR}} - \rho_{\text{Red}})/(\rho_{\text{NIR}} + \rho_{\text{Red}}) \tag{20}$$

and ρ_{NIR} and ρ_{Red} are the measured reflectance factors of the surface in the near-infrared (NIR) and red spectrum, respectively. They suggest that this relationship is related to the amount of available energy partitioned into λE, which is driven by variation in transpiration or evaporative cooling. Hope et al. (1986) showed theoretically that with VI and T_{rad} one can extract canopy resistance. However, this assumes complete canopy cover, which does not usually exist in most natural land surfaces.

Nemani and Running (1989) used an ecological model for forested regions and observed a nonlinear relationship between the slope of the NDVI–T_{rad} curve and the canopy resistance. Goward and Hope (1989) also proposed that the slope was a measure of the surface resistance. These approaches will be difficult to apply to most landscapes with partial canopy cover since variability in fractional cover and surface soil moisture cause significant scatter in the VI/T_{rad} relationship. Furthermore, studies suggest that the relationship between surface resistance and the NDVI/T_{rad} slope will vary significantly with vegetation type. Nemani et al. (1993) showed that the NDVI/T_{rad} slope responded to changes in water status of forested areas, but not of the grasslands. The variability in slope for the grasslands appeared to be mainly caused by variation in fractional cover rather than in λE. Smith and Choudhury (1991) used a coupled dual-source soil–vegetation model to show that the NDVI/T_{rad} slope largely depended on whether the drying soil surface is the source of the decline in λE or whether it was the vegetation. They also observed that the linear relationship between NDVI and T_{rad} did not exist for forests but only for agricultural and native pastures.

Others have used an energy balance model for computing spatially distributed fluxes from the variability within the NDVI–T_{rad} plot from a single scene (Price, 1990). Price (1990) used NDVI to estimate the fraction of a pixel covered by vegetation. From the NDVI/T_{rad} plot Price (1990) showed how one could derive bare soil and vegetation temperatures and, with enough spatial variation in surface moisture, estimate daily λE for the limits of full cover vegetation, dry and wet bare soils.

Following Price (1990), Carlson et al. (1990, 1994) combined an ABL model with a SVAT for mapping surface soil moisture, vegetation cover, and surface fluxes. Model simulations were run for two conditions: 100% vegetative cover with the maximum NDVI being known a priori, and with bare soil conditions knowing the

minimum NDVI. Using ancillary data (including a morning atmospheric sounding, vegetation and soil-type information) root-zone and surface soil moisture were varied, respectively, until the modeled and measured T_{rad} were closely matched for both cases, and fractional vegetated cover and surface soil moisture were derived. Further refinements to this technique have been developed by Gillies and Carlson (1995) for potential incorporation into climate models. Comparisons between modeled-derived fluxes and observations have been made recently by Gillies et al. (1997) using high-resolution aircraft-based remote-sensing measurements from a grassland ecosystem during FIFE and Monsoon'90. Approximately 90% of the variance in the fluxes was explained by the model.

In a related approach, Moran et al. (1994) defined theoretical boundaries in the SAVI/$(T_{rad} - T_a)$ two-dimensional space using the Penman–Monteith equation, where SAVI is the soil-adjusted vegetation index proposed by Huete (1988). The boundaries define a trapezoid, which has at the upper two corners unstressed and stressed 100% vegetated cover and at the lower two corners wet and dry bare soil conditions. To calculate the vertices of the trapezoid, measurements of R_n, vapor pressure, T_a, and wind speed are required as well as vegetation-specific parameters; these include maximum and minimum SAVI for the full-cover and bare soil case, maximum leaf area index, and maximum and minimum stomatal resistance. Moran et al. (1994) analyzed and discussed several of the assumptions underlying the model, especially those concerning the linearity between variations in canopy–air temperature and soil–air temperatures and transpiration and evaporation. Information about λE rates is derived from the location of the SAVI/$(T_{rad} - T_a)$ measurements within the date and time-specific trapezoid. This approach permits the technique to be used for both heterogeneous and uniform areas and thus does not require having a range of NDVI and surface temperature in the scene of interest as required by Carlson et al. (1990) and Price (1990). Moran et al. (1994) compared the method for estimating relative rates of λE with observations over agricultural fields and showed it could be used for irrigation scheduling purposes. More recently, Moran et al. (1996) showed that the technique had potential for computing λE over natural grassland ecosystems.

5 SYNTHESIS

In this chapter, numerous methods were reviewed for using remote sensing to estimate λE. Based on a similar review conducted by Kustas and Norman (1996), a series of issues were identified as important for remote sensing of λE from measurements, modeling studies, and theoretical considerations. A slightly revised list of these issues is included here:

1. T_{rad} is not equal to T_{aero}.
2. Most models are sensitive to errors in $T_{aero} - T_a$ and u, yet the measurement of T_a and u at the time and location of the T_{rad} observation is not typically available.

3. T_{rad} dependence on view angle cannot generally be neglected because differences in vegetation and soil temperatures can be significant depending on soil moisture conditions.
4. Thermal emissivity is only known approximately on the pixel scale.
5. Atmospheric corrections and satellite calibrations contribute significant errors in the measurements of $\rho_{\Delta\lambda}$ and T_{rad} that are not always adequately known.
6. Remote observations are instantaneous, while integrated fluxes are desired on hourly, daily, or longer time scales.
7. Satellites with larger pixel sizes (1 to 4 km) can provide sufficiently frequent observations in time (i.e., *GOES*), but may have uncertainties related to the averaging over heterogeneous subpixel areas.
8. Continuous (hourly or daily) surface flux estimates are most useful, and clouds cause remote observations to be intermittent.

Kustas and Norman (1996) provided a representative list of models using remote observations to estimate λE and attempted to characterize which of the above eight issues each of these models addressed. None of the models address *all* the important issues at the present time, but several of the models address some of the important issues (1, 3, 4, and 6). Fewer models addressed the most critical issues of spatially distributed meteorological data and atmospheric correction of satellite image data (2 and 5). Related to issue 2, meteorological data acquired at a time or location other than that of the T_{rad} or VI observation can cause substantial error in the estimate of λE. Moran and Jackson (1991) reported that errors in extrapolation of T_a greater than 1°C were unacceptable for estimation of λE using the energy balance approach. They also reported that measurements of T_a measured at 2 m height over adjacent fields of bare soil and lush vegetation differed by up to 3°C at midday. Similarly disturbing results have been reported for wind speed estimation. Rahman (1996) compared a wind speed map constructed by simple interpolation of u values from local weather stations with a map of wind speed derived from the Regional Atmospheric Modeling System (RAMS; Pielke et al., 1992) that accounted for topographic effects. The RAMS-derived map of u was a substantial improvement over the simple interpolation because it accounted for the relatively strong winds in the passes between mountain ranges and relatively light winds in the lee of the ranges.

Related to issue 5, accounting for the attenuation of the radiances received by satellite-based sensors is not a trivial matter (Kaufman, 1989; Price, 1989). In correcting thermal-infrared data, whether using radiative transfer models or split-window techniques, the uncertainty is 1 to 3°C over land surfaces (Becker and Li, 1990; Perry and Moran, 1994). Model sensitivity to such an uncertainty in T_{rad} can be significant, especially over large vegetation where errors can be $\sim 100\,W/m^2$ for hourly to daily time scales (Norman et al., 1995a). However, the $150\,W/m^2$ uncertainty in estimating sensible heat flux from radiometric surface temperature observations suggested by Sellers et al. (1995b) is in many cases two to three times larger than errors reported by other researchers (Choudhury, 1994). All the methods reviewed in this chapter are based on the assumption that accurate remotely sensed estimates of surface reflectance, temperature, and backscatter will be readily

available. At this time, they are not. A primary challenge will be to improve the accuracy and consistency of remotely sensed information with an insight into the accuracy requirements of operational models and algorithms.

None of the models explicitly addressed the issue of subpixel averaging, often termed *aggregation* (issue 7). Aggregation refers to spatial averaging of some heterogeneous surface variable to obtain an effective value representative of an area. In an assessment of the state of the art in aggregation research, Michaud and Shuttleworth (1997) concluded that, over flat terrain, simple aggregation rules applied to surface properties could result in simulated values of λE within 10% of fluxes from models with full representation of heterogeneity. Furthermore, they concluded that aggregation rules for vegetation characteristics were relatively straightforward in the case of patch-scale heterogeneity (variability of 100 to 1000 m). However, mesoscale heterogeneity (10 to 100 km) in surface cover will need to be addressed through more complicated types of parameterization and, in mountainous terrain, the influence of topography on near-surface meteorology must be considered. In an aggregation study related to the use of remote-sensing data for energy balance evaluation, Moran et al. (1997a) found that aggregation of remotely sensed measurements in sparse canopies could be accomplished with little error (such as aggregation of T_{rad} from 1 m^2 to 1 km^2) but not others (such as aggregation of H to 1 km^2). Kustas and Humes (1996) applied the Norman et al. (1995b) dual-source model for computing basin-scale fluxes with T_{rad} at 120-, 1000-, and ~8000-m pixel resolution over a semiarid rangeland landscape. They found minor changes in the fluxes aggregated from the different resolutions. Sellers et al. (1995a) investigated the impact of spatial variation in topography, vegetative cover, and soil moisture on area-averaged fluxes simulated by a SVAT model over a 2×15 km domain. They found simple averages of these parameters introduced minor errors in the SVAT simulations of the area-averaged fluxes. Still, other studies (Crosson et al., 1993; Sellers et al., 1992) suggest that issue 7 may be a significant problem at the 1-km scale but may average out at the 10-km scale (Norman and Divakarla, 1995).

None of the current models address the issue of continuous surface fluxes even with clouds, but studies are in progress to combine the thermal infrared remote-sensing approaches discussed in this chapter with mesoscale models and with a simplified land–atmosphere exchange model (Anderson et al., 2000). If issues 1 to 7 are addressed adequately, issue 8 will not limit remote estimation of regional λE fluxes.

6 CONCLUDING REMARKS

All the methods and models reviewed in this chapter have potential for operational evaluation of the spatial distribution of evaporation for agricultural and hydrological applications. Toward that goal, relatively simple methods using one-time-of-day remote sensing observations for quantifying daily ET have been applied operationally (Seguin et al., 1989, 1991). However, for many regions of Earth's land surface, meteorological data (primarily wind speed and air temperature) essential for driving

model computations are not available. Approaches using remotely sensed data for estimating the variation of these quantities are being developed and tested (Bastiaanssen et al., 1998; Gao et al., 1998). How reliable the algorithms are for different climatic regimes needs to be evaluated. For air temperature, another approach is in the utilization of radiometric temperature observations from significantly different view angles in a dual-source model (Kustas and Norman, 1997). SVAT models using remote-sensing observations and linked to operational climate and hydrologic models (Ottlé and Vidal-Madjar, 1994; Gillies and Carlson, 1995; Mecikalski et al., 1999; Nouvellon et al., 2001) probably have the greatest potential for operational, regional application. This is because both the surface boundary conditions and atmospheric variables are simulated over time. For heterogenous and mountainous landscapes, further work should be focused on the development of robust aggregation techniques (e.g., Shuttleworth, 1998).

One of the greatest obstacles to the assimilation of remotely sensed information in physical models has been the inherent limitations of currently available sensors. Satellite-based sensors have the advantages of good geometric and radiometric integrity; the disadvantages include fixed spectral bands that may be inappropriate for a given application, spatial resolutions too coarse or too fine for the application, long time periods between image acquisition and delivery to user, and inadequate repeat coverage due to sensor or weather limitations. With the exception of the limitations due to weather, many of the existing limitations may be resolved with the newly launched *Terra*, *Landsat-7*, and *Space Imaging* satellites (Table 2).

Regarding the effects of clouds on image acquisitions, more work should be directed toward utilizing microwave remote sensing, which has some critical advantages over the use of optical data, including little atmospheric attenuation, cloud penetration, high spatial resolution, and day/night acquisitions. Microwave data have been used to derive soil moisture and other vegetation properties (Jackson et al., 1995; Moran et al., 1997b). Microwave data have also been used for estimating the partitioning of available energy into H and λE, for estimating soil evaporation, and in determining soil surface temperatures (Kustas et al., 1993b; Chanzy and Kustas, 1995; Troufleau et al., 1994). More recently, the dual-source model of Norman et al. (1995b) was revised to use remotely sensed near-surface moisture from a passive microwave sensor for estimating the soil surface energy balance (Kustas et al., 1998). With remotely sensed images of near-surface soil moisture, land cover classification and LAI, the model was applied over a semiarid area in southern Arizona. Comparison of model-predicted fluxes simulated over the daytime period with ground observations showed good results, with 15% differences in evaporation estimates, on average. It is also shown that it may be possible to simulate the daytime fluxes with only a single microwave observation.

The development of methods for combining microwave and optical data with SVAT schemes will likely produce the greatest advancement in the quantification of spatially distributed evaporation. This requires collection of remote-sensing data in concert with ground observations as part of large-scale field projects conducted in different climatic regions. This is a critical part in the further development and

validation of model algorithms. Thus the conventional approaches for estimating evaporation outlined in this chapter play a key role in this effort.

ACKNOWLEDGMENT

The authors acknowledge funding support from NASA under grant numbers NAGW-4138, NASA-S-41396-F, and NASA NAGW-2425 through the NASA Earth Science Enterprise. The authors would like to thank Drs. Bruce Goff and Martha Anderson for their comments on an early version of this chapter.

REFERENCES

Anderson, M. C., J. M. Norman, G. R. Diak, W. P. Kustas, and J. R. Mecikalski, A two-source time-integrated model for estimating surface fluxes from thermal infrared satellite observations, *Remote Sensing Environ.*, *60*, 195–216, 1997.

Anderson, M. C., J. M. Norman, T. P. Meyers and G. R. Diak, An analytical model for estimating canopy transpiration and carbon assimilation fluxes based on light-use efficiency, *Agric. For. Meteor.*, *100*, 265–289, 2000.

Andre, J. C., J. P. Goutorbe, and A. Perrier, HAPEX-MOBILHY: A hydrologic atmospheric experiment for the study of water budget and evaporation flux at the climatic scale, *Bull. Am. Meteor. Soc.*, *67*, 138–144, 1986.

Bastiaanssen, W. G. M., R. A. Feddes, and A. A. M. Holtslag, A remote sensing surface energy balance algorithm for land (SEBAL) Part 1: Formulation, *J. Hydrol.*, 212–213: 198–2121998.

Becker, F., and Z. L. Li, Towards a local split window method over land surfaces, *Int. J. Remote Sensing*, *11*, 369–393, 1990.

Beljaars, A. C. M., and A. A. M. Holtslag, Flux parameterization over land surfaces for atmospheric models, *J. Appl. Meteor.*, *30*, 327–341, 1991.

Bougeault, P., J. Noilhan, P. Lacarrer, and P. Mascart, An experiment with an advanced surface parameterization in a mesobeta-scale model. Part I: Implementation. *Monthly Weather Rev.*, *119*, 2358–237, 1991.

Bowen, I. S., The ratio of heat losses by conduction and by evaporation from any water surface, *Phys. Rev.*, *27*, 779–789, 1926.

Brunet, Y., M. Nunez, and J.-P. Lagouarde, A simple method for estimating regional evapotranspiration from infrared surface temperature data, *ISPRS J. Photogram. Remote Sensing*, *46*, 311–327, 1991.

Brutsaert, W., *Evaporation into the Atmosphere*, Reidel, Dordrecht, 1982.

Brutsaert, W., and J. A. Mawdsley, Applicability of planetary boundary layer theory to calculate regional evapotranspiration, *Water Resour. Res.*, *12*, 852–858, 1976.

Brutsaert, W., and M. Sugita, A bulk similarity approach in the atmospheric boundary layer using radiometric skin temperature to determine regional fluxes, *Boundary-Layer Meteor.*, *55*, 1–23, 1991.

Brutsaert, W., and M. Sugita, Application of self-preservation in the diurnal evolution of the surface energy budget to determine daily evaporation, *J. Geophys. Res.*, *97*(D17), 18377–18382, 1992.

Brutsaert, W., A. Y. Hsu, and T. J. Schmugge, Parameterization of surface heat fluxes above a forest with satellite thermal sensing and boundary layer soundings, *J. Appl. Meteor.*, *32*, 909–917, 1993.

Camillo, P., Estimating soil surface temperatures from profile temperature and flux measurements. *Soil Sci.*, *148*, 233–243, 1989.

Camillo, P. J., R. J. Gurney, and T. J. Schmugge, A soil and atmospheric boundary layer model for evapotranspiration and soil moisture studies, *Water Resour. Res.*, *19*, 371–380, 1983.

Campbell, G. S., *Soil Physics with Basic*, Elsevier, New York, 1985.

Carlson, T. N., and M. J. Buffum, On estimating total daily evapotranspiration from remote surface measurements, *Remote Sensing Environ.*, *29*, 197–207, 1989.

Carlson, T. N., J. K. Dodd, S. G. Benjamin, and J. N. Cooper, Satellite estimation of the surface energy balance, moisture availability and thermal inertial, *J. Appl. Meteor*, *20*, 67–87, 1981.

Carlson, T. N., E. M. Perry, and T. J. Schmugge, Remote estimation of soil moisture availability and fractional vegetation cover for agricultural fields, *Agric. For. Meteor.*, *52*, 45–69, 1990.

Carlson, T. N., R. R. Gillies, and E. M. Perry, A method to make use of thermal infrared temperature and NDVI measurements to infer soil water content and fractional vegetation cover, *Remote Sensing Rev.*, *52*, 45–59, 1994.

Carlson, T. N., W. J. Capehart, and R. R. Gillies, A new look at the simplified method for remote sensing of daily evapotranspiration, *Remote Sensing Environ.*, *54*, 161–167, 1995.

Chanzy, A., and W. P. Kustas, Evaporation monitoring over land surface using microwave radiometry, in B. J. Choudhury, Y. H. Kerr, E. G. Njoku, and P. Pampaloni (Eds.), *ESA/NASA International Workshop*, VSP, Utrecht, 1995, pp. 531–550.

Chehbouni, A., D. Lo Seen, E. G. Njoku, and B. M. Monteney, Examination of the difference between radiative and aerodynamic surface temperatures over sparsely vegetated surfaces, *Remote Sensing Environ.*, *58*, 177–186, 1996.

Choudhury, B. J., Synergism of multispectral satellite observations for estimating regional land surface evaporation, *Remote Sensing Environ.*, *49*, 264–274, 1994.

Choudhury, B. J., J. R. Reginato, and S. B. Idso, An analysis of infrared temperature observations over wheat and calculation of latent heat flux, *Agric. Forest Meteor.*, *37*, 75–88, 1986.

Choudhury, B. J., S. B. Idso, and J. R. Reginato, Analysis of an empirical model for soil heat flux under a growing wheat crop for estimating evaporation by an infrared-temperature based energy balance equation, *Agric. Forest Meteor.*, *39*, 283–297, 1987.

Choudhury, B. J., N. U. Ahmed, S. B. Idso, R. J. Reginato, and C. S. T. Daughtry, Relations between evaporation coefficients and vegetation indices studied by model simulations, *Remote Sensing Environ.*, *50*, 1–17, 1994.

Clothier, B. E., K. L. Clawson, P. J. Pinter, Jr., M. S. Moran, R. J. Reginato, and R. D. Jackson, Estimation of soil heat flux from net radiation during the growth of alfalfa, *Agric. Forest Meteor.*, *37*, 319–329, 1986.

Crosson, W. L., E. A. Smith, and H. J. Cooper, Estimation of surface heat and moisture fluxes over a prairie grassland. 4: Impact of satellite remote sensing of slow canopy variables on performance of a hybrid biosphere model, *J. Geophys. Res.*, *98*(D3), 4979–4999, 1993.

Darnell, W. L., W. F. Staylor, S. K. Gupta, N. A. Ritchey, and A. C. Wilber, Seasonal variation of surface radiation budget derived from International Satellite Cloud Climatology Project C1 data, *J. Geophys. Res.*, *97*, 15741–15760, 1992.

Daughtry, C. S. T., W. P. Kustas, M. S. Moran, P. J. Pinter, Jr., R. D. Jackson, P. W. Brown, W. D. Nichols, and L. W. Gay, Spectral estimates of net radiation and soil heat flux, *Remote Sensing Environ.*, *32*, 111–124, 1990.

Diak, G. R., Evaluation of heat flux, moisture flux and aerodynamic roughness at the land surface from knowledge of the PBL height and satellite-derived skin temperatures, *Agric. Forest Meteor.*, *52*, 181–198, 1990.

Diak, G. R., and M. A. Whipple, Improvements to models and methods for evaluating the land-surface energy balance and "effective" roughness using radiosonde reports and satellite-measured "skin" temperatures, *Agric. Forest Meteor.*, *63*, 189–218, 1993.

Diak, G. R., C. J. Scheuer, M. S. Whipple, and W. L. Smith, Remote sensing of land-surface energy balance using data from the high-resolution interferometer sounder (HIS): A simulation study, *Remote Sensing Environ.*, *48*, 106–118, 1994.

Diak, G. R., R. M. Rabin, K. P. Gallo, and C. M. U. Neale, Regional-scale comparisons of NDVI, soil moisture indices from surface and microwave data and surface energy budgets evaluated from satellite and in-situ data, *Remote Sensing Rev.*, *12*, 355–382, 1995.

Gao, W., R. L. Coultier, B. M. Lesht, J. Qui, and M. L. Wesely, Estimating clear-sky regional surface fluxes in the southern Great Plains atmospheric radiation measurement site with ground measurements and satellite observations, *J. Appl. Meteorol.*, *37*, 5–22, 1998.

Gay, L. W., and R. J. Greenberg, The AZET battery-powered Bowen ratio system, in *Proceedings of the 17th Conf. on Agric. and Forest. Meteorol.*, 21–23 May, 1985, Scottsdale, AZ. American Meteorological Society, Boston, MA, 1985, pp. 181–182.

Gillies, R. R., and T. N. Carlson, Thermal remote sensing of surface soil water content with partial vegetation cover for incorporation into climate models, *J. Appl. Meteor.*, *34*, 745–756, 1995.

Gillies, R. R., T. N. Carlson, J. Cui, W. P. Kustas, and K. S. Humes, Verification of the "triangle" method for obtaining surface soil water content and energy fluxes from remote measurements of Normalized Difference Vegetation Index (NDVI) and surface radiant temperature, *Int. J. Remote Sensing*, *18*, 3145–3166, 1997.

Goudriaan, J., *Crop Micrometeorology: A Simulation Study*, Center for Agric., 1977.

Goward, S. N., and A. S. Hope, Evapotranspiration from combined reflected solar and emitted terrestrial radiation: Preliminary FIFE results from AVHRR data, *Adv. Space Res.*, *9*, 239–249, 1989.

Goward, S., G. D. Cruickshanks, and A. Hope, Observed relation between thermal emission and reflected spectral radiance of a complex vegetated landscape, *Remote Sensing Environ.*, *18*, 137–146, 1985.

Goward, S. N., R. H. Waring, D. G. Dye, and J. Yang, Ecological remote sensing at OTTER: Satellite macroscale observations, *Ecol. Appl.*, *4*, 322–343, 1994.

Hall, F. G., K. F. Huemmrich, S. J. Geotz, P. J. Sellers, and J. E. Nickerson, Satellite remote sensing of surface energy balance: Success, failures and unresolved issues in FIFE, *J. Geophys. Res.*, *97*(D17), 19061–19090, 1992.

Harshvardhan, R. D. A., and D. A. Dazlich, Relationship between the longwave cloud radiative forcing at the surface and the top of the atmosphere, *J. Clim.*, *3*, 1435–1443, 1990.

Holmes, J. W., Measuring evapotranspiration by hydrological methods, *Agric. Water Mgmt.*, *8*, 29–40, 1984.

Hope, A. S., and T. P. McDowell, The relationship between surface temperature and a spectral vegetation index of a tallgrass prairie: Effects of burning and other landscape controls, *Int. J. Remote Sensing*, *13*, 2849–2863, 1992.

Hope, A. S., D. E. Petzold, S. N. Goward, and R. M. Ragan, Simulated relationships between spectral reflectance, thermal emissions, and evapotranspiration of a soybean canopy, *Water Resour. Bull.*, *22*, 1011–1019, 1986.

Howell, T. A., R. L. McCormick, and C. J. Phene, Design and installation of large weighing lysimeters, *Trans. Am. Soc. Agric. Eng.*, *28*, 106–112, 117, 1985.

Huband, N. D. S., and J. L. Monteith, Radiative surface temperature and energy balance of a wheat canopy. Part I: Comparison of radiative and aerodynamic canopy temperature, *Bound.-Layer Meteor.*, *36*, 1–17, 1986.

Huete, A. R., A soil-adjusted vegetation index (SAVI), *Remote Sensing Environ.*, *27*, 47–57, 1988.

Jackson, R. D., and A. R. Huete, Interpreting vegetation indices, *Prev. Vet. Med.* 11, 185–200, 1991.

Jackson, R. D., R. J. Reginato, and S. B. Idso, Wheat canopy temperature: A practical tool for evaluating water requirements, *Water Resour. Res.*, *13*, 651–656, 1977.

Jackson, R. D., J. L. Hatfield, R. J. Reginato, S. B. Idso, and P. J. Pinter, Jr., Estimates of daily evapotranspiration from one time of day measurements, *Agric. Water Mgmt.*, *7*, 351–362, 1983.

Jackson, R. D., M. S. Moran, L. W. Gay, and L. H. Raymond, Evaluating evaporation from field crops using airborne radiometry and ground-based meteorological data, *Irrig. Sci.*, *8*, 81–90, 1987.

Jackson, T. J., P. E. O'Neill, W. P. Kustas, E. Bennett, and C. T. Swift, Passive microwave observation of diurnal soil moisture at 1.4 and 2.65 GHz, in *Proceedings of the 1995 International Geoscience and Remote Sensing Symposium*, T. I. Stein (Ed.), Vol. I, Institute of Electrical and Electronics Engineers, New York, pp. 492–494, 1995.

Jensen, M. E., and J. L. Wright, The role of evapotranspiration models in irrigation scheduling, *Trans. Am. Soc. Agric. Eng.*, *21*, 82–87, 1978.

Jensen, M. E., R. D. Burman, and R. G. Allen (Eds.), *Evapotranspiration and Irrigation Water Requirements: A Manual*, No. 70, Am. Soc. Civil. Eng. (ASCE) New York, NY, 1989.

Kanemasu, E. T., M. L. Wesely, B. B. Hicks, and J. L. Heilman, Techniques for calculating energy and mass fluxes, in B. J. Barfield and J. F. Gerber (Eds.), *Modification of the Aerial Environment of Crops*, American Society of Agricultural Engineers, St. Joseph, MI, 1979, pp. 156–182.

Kaufman, Y. J. The atmospheric effect on remote sensing and its corrections, in G. Asrar (Ed.), *Theory and Applications of Optical Remote Sensing*, Wiley, New York, 1989, pp. 336–428.

Kohsiek, W., H. A. R. De Bruin, H. The, and B. van den Hurk, Estimation of the sensible heat flux of semi-arid area using surface radiative temperature measurements, *Boundary-Layer Meteor.*, *63*, 213–230, 1993.

Kustas, W. P., Estimates of evapotranspiration with a one- and two-layer model of heat transfer over partial canopy cover, *J. Appl. Meteor.*, *29*, 704–715, 1990.

Kustas, W. P., and D. C. Goodrich, Preface to Monsoon '90 Special Section, *Water Resour. Res.*, *30*, 1211–1225, 1994.

Kustas, W. P., and K. S. Humes, Variations in the surface energy balance for a semi-arid rangeland using remotely sensed data at different spatial resolutions, in J. B. Stewart, E. T. R. Engman, A. Feddes, and Y. Kerr (Eds.), *The Scaling Issue in Hydrology*, Wiley, New York, 1996, pp. 127–145.

Kustas, W. P., and J. M. Norman, Use of remote sensing for evapotranspiration monitoring over land surfaces, *Hydrol. Sci. J. Sci. Hydrol.*, *41*, 495–516, 1996.

Kustas, W. P., and J. M. Norman, A two-source approach for estimating turbulent fluxes using multiple angle thermal infrared observations, *Water Resour. Res.*, *33*, 1495–1508, 1997.

Kustas, W. P., B. J. Choudhury, M. S. Moran, R. J. Reginato, R. D. Jackson, L. W. Gay, an H. L. Weaver, Determination of sensible heat flux over sparse canopy using thermal infrared data, *Agric. Forest Meteor.*, *44*, 197–216, 1989.

Kustas, W. P., C. S. T. Daughtry, and P. J. van Oevelen, Analytical treatment of the relationships between soil heat flux/net radiation ratio and vegetation indices, *Remote Sensing Environ.*, *46*, 319–330, 1993a.

Kustas, W. P., T. J. Schmugge, K. S. Humes, T. J. Jackson, R. Parry, M. A. Weltz, and M. S. Moran, Relationships between evaporative fraction and remotely sensed vegetation index and microwave brightness temperature for semiarid rangelands, *J. Appl. Meteor.*, *32*, 1781–1790, 1993b.

Kustas, W. P., E. M. Perry, P. C. Doraiswamy, and M. S. Moran, Using satellite remote sensing to extrapolate evapotranspiration estimates in time and space over a semiarid rangeland basin, *Remote Sensing Environ.*, *49*, 275–286, 1994a.

Kustas, W. P., R. T. Pinker, T. J. Schmugge, and K. S. Humes, Daytime net radiation estimated for a semiarid rangeland basin from remotely sensed data, *Agric. Forest Meteor.*, *71*, 337–357, 1994b.

Kustas, W. P., K. S. Humes, J. M. Norman, and M. S. Moran, Single- and dual-source modeling of surface energy fluxes with radiometric surface temperature, *J. Appl. Meteor.*, *35*, 110–121, 1996.

Kustas, W. P., X. Zhan, and T. J. Schmugge, Combining optical and microwave remote sensing for mapping energy fluxes in a semiarid watershed, *Remote Sensing Environ.*, 64, 116–131, 1998.

Lagouarde, J.-P., Use of NOAA AVHRR data combined with an agrometeorological model for evaporation mapping, *Int. J. Remote Sensing*, *12*, 1853–1864, 1991.

Lagouarde, J.-P., and K. J. McAneney, Daily sensible heat flux estimation from a single measurement of surface temperature and maximum air temperature, *Boundary-Layer Meteor.*, *59*, 341–362, 1992.

Lhomme, J.-P., B. Monteny, and M. Amadou, Estimating sensible heat flux from radiometric temperature over sparse millet, *Agric. Forest Meteor.*, *68*, 77–91, 1994.

McNaughton, K. G., and T. W. Spriggs, A mixed-layer model for regional evaporation, *Boundary-Layer Meteor.*, *34*, 243–262, 1986.

McNaughton, K. G., and B. J. J. M. Van den Hurk, A "Lagrangian" revision of the resistors in the two-layer model for calculating the energy budget of a plant canopy, *Boundary-Layer Meteor.*, *74*, 262–288, 1995.

Mecikalski, J. R., G. R. Diak, M. C. Anderson and J. M. Norman, 1999. Estimating fluxes on continental scales using remotely-sensed data in an atmospheric-land exchange model. *J. Appl. Meterol.*, *38*, 1352–1369.

Michaud, J. D., and W. J. Shuttleworth, Executive summary of the Tucson Aggregation Workshop, *J. Hydrol.*, *190*, 176–181, 1997.

Monteith, J. L., Evaporation and environment, *Symp. Soc. Exp. Biol.*, *19*, 205–234, 1965.

Monteith, J. L., Evaporation and surface temperature, *Q. J. R. Meteor. Soc.*, *107*, 1–27, 1981.

Moran, M. S., and R. D. Jackson, Assessing the spatial distribution of evapotranspiration using remotely sensed inputs, *J. Environ. Q.*, *20*, 725–737, 1991.

Moran, M. S., R. D. Jackson, L. H. Raymond, L. W. Gay, and P. N. Slater, Mapping surface energy balance components by combining LANDSAT Thematic Mapper and ground-based meteorological data, *Remote Sensing Environ.*, *30*, 77–87, 1989.

Moran, M. S., T. R. Clarke, Y. Inoue, and A. Vidal, Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index, *Remote Sensing Environ.*, *49*, 246–263, 1994.

Moran, M. S., A. F. Rahman, J. C. Washburne, D. C. Goodrich, M. A. Weltz, and W. P. Kustas, Combining the Penman–Monteith equation with measurements of surface temperature and reflectance to estimate evaporation rates of semiarid grassland, *Agric. Forest Meteor.*, *80*, 87–109, 1996.

Moran, M. S., K. S. Humes, and P. J. Pinter, Jr., The scaling characteristics of remotely sensed variables for sparsely-vegetated heterogeneous landscapes, *J. Hydrol.*, *190*, 338–363, 1997a.

Moran, M. S., A. Vidal, D. Troufleau, J. Qi, T. R. Clarke, P. J. Pinter, Jr., T. Mitchell, Y. Inoue, and C. M. U. Neale, Combining multifrequency microwave and optical data for farm management, *Remote Sensing Environ.*, *61*, 96–109, 1997b.

Nemani, R. R., and S. W. Running, Estimation of regional surface resistance to evapotranspiration from NDVI and thermal-IR AVHRR data, *J. Appl. Meteor.*, *28*, 276–284, 1989.

Nemani, R., L. Pierce, S. Running, and S. Goward, Developing satellite derived estimates of surface moisture status, *J. Appl. Meteor.*, *32*, 548–557, 1993.

Nie, D., E. T. Kanemasu, L. J. Fritschen, H. L. Weaver, E. A. Smith, S. B. Verma, R. T. Field, W. P. Kustas, and J. B. Stewart, An intercomparison of surface energy flux measurement systems during FIFE 1987, *J. Geophys. Res.*, *97*(D17), 18715–18724, 1992.

Nieuwenhuis, G. J. A., E. A. Schmidt, and H. A. M. Tunnissen, Estimation of regional evapotranspiration of arable crops from thermal infrared images, *Int. J. Remote Sensing*, *6*, 1319–1334, 1985.

Norman, J. M., and F. Becker, Terminology in thermal infrared remote sensing of natural surfaces, *Remote Sensing Rev.*, *12*, 159–173, 1995.

Norman, J. M., and M. Divakarla, Scaling carbon, water and energy fluxes from 30 m to 15 km, in *Agronomy Abstracts*, American Society of Agronomy Madison, WI, 1995.

Norman, J. M., M. Divakarla, and N. S. Goel, Algorithms for extracting information from remote thermal-IR observations of the earth's surface, *Remote Sensing Environ.*, *51*, 157–168, 1995a.

Norman, J. M., W. P. Kustas, and K. S. Humes, A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature, *Agric. Forest Meteor.*, *77*, 263–293, 1995b.

Nouvellon, Y., M. S. Moran, D. Lo Seen, R. B. Bryant, W. Ni, A. Begue, A. G. Chehbouni, W. E. Emmerich, P. Heilman and J. Qi, Coupling a grassland ecosystem model with Landsat imagery for a 10-year simulation of carbon and water budgets, *Rem. Sens. Env. 78*:131–149, 2001.

Ottlé, C., and D. Vidal-Madjar, Assimilation of soil moisture inferred from infrared remote sensing in a hydrological model over the HAPEX-MOBILHY region, *J. Hydrol., 158*, 241–264, 1994.

Owe, M., and A. A. van de Griend, Daily surface moisture model for large area semiarid land application with limited climate data, *J. Hydrol., 121*, 119–132, 1990.

Penman, H. L., Natural evaporation from open water, bare soil and grass, *Proc. R. Soc. A., 193*, 120–145, 1948.

Penman, H. L., The physical bases of irrigation control, *Rep. 13th Int. Hort. Cong., 2*, 913–923, 1953.

Perry, E. M., and M. S. Moran, An evaluation of atmospheric corrections of radiometric surface temperatures for a semiarid rangeland watershed, *Water Resour. Res., 30*, 1261–1269, 1994.

Pielke, R. A., W. R. Cotton, R. L. Walko, C. J. Tremback, W. A. Lyons, L. D. Grasso, M. E. Nicholls, M. D. Moran, D. A. Wesley, T. J. Lee, and J. H. Copeland, A comprehensive meteorological modeling system: RAMS, *Meteor. Atmos. Phys., 49*, 69–91, 1992.

Pinker, R. T., and J. D. Tarpley, The relationship between the planetary and surface net radiation: An update, *J. Appl. Meteor., 27*, 957–964, 1988.

Pinker, R. T., W. P. Kustas, I. Laszlo, M. S. Moran, and A. R. Huete, Satellite surface radiation budgets on basin scale in semi-arid regions, *Water Resour. Res., 30*, 1375–1386, 1994.

Pinker, R. T., R. Frovin, and Z. Li, A review of satellite methods to derive surface shortwave irradiance, *Remote Sensing Environ., 51*, 108–124, 1995.

Prata, A. J., Land surface temperatures derived from the AVHRR and ATSR I: Theory, *J. Geophys. Res., 89*(D9): 16689–16702, 1993.

Prata, A. J., R. P. Cechet, I. J. Barton, and D. T. Llewellyn-Jones, The along track scanning radiometer for ERS-1-scan geometry and data simulation, *IEEE Trans. Geosci. Remote Sensing, 28*, 3–13, 1990.

Prévot, L., K. T. Brunet, U. Paw, and B. Seguin, Canopy modelling for estimating sensible heat flux from thermal infrared measurements, in *Proceedings of the Workshop on Thermal Remote Sensing of the Energy and Water Balance over Vegetation in Conjunction with Other Sensors*, Cemagref-Engref, Montpellier, France, 1994, pp. 17–26.

Price, J. C., The potential of remotely sensed thermal infrared data to infer surface soil moisture and evaporation, *Water Resour. Res., 16*, 787–795, 1980.

Price, J. C., Estimation of regional scale evaporation through analysis of satellite thermal-infrared data, *IEEE Trans. Geosci. Remote Sensing, GE-20*, 286–292, 1982.

Price, J. C., Quantitative aspects of remote sensing in the thermal infrared, in G. Asrar (Ed.), *Theory and Applications of Optical Remote Sensing*, Wiley, New York, 1989, pp. 578–603.

Price, J. C., Using spatial context in satellite data to infer regional scale evapotranspiration, *IEEE Trans. Geosci. Remote Sensing, GE-28*, 940–948, 1990.

Priestley, C. H. B., and R. J. Taylor, On the assessment of surface heat flux and evaporation using large scale parameters, *Monthly Weather Rev., 100*, 81–102, 1972.

Rahman, A. F., Monitoring regional-scale surface hydrologic processes using satellite remote sensing, Ph.D. dissertation, University of Arizona, Department of Soil and Water Science, Tucson, AZ, 1996.

Reicosky, D. C., A research tool for evapotranspiration measurements for model validation and irrigation scheduling, in *Irrigation Scheduling for Water and Energy Conservation in the 80's, Proc. of the Am. Soc. of Agric. Engineers Irrig. Scheduling Conf.*, December 1981, ASAE, Chicago, IL, 1981, pp. 18–26.

Rijtema, P. R., An analysis of actual evapotranspiration, *Agric. Res. Rep.*, *659*, 1–107, 1965.

Robins, J. R., W. O. Pruitt, and W. H. Gardner, Unsaturated flow of water in field soils and its effect on soil moisture investigations, *Soil. Sci. Soc. Am. Proc.*, *18*, 344–347, 1954.

Sauer, T. J., J. M. Norman, C. B. Tanner, and T. B. Wilson, Measurement of heat and vapor transfer coefficients at the soil surface beneath a maize canopy using source plates, *Agric. Forest Meteor.*, *75*, 161–189, 1995.

Seguin, B., and B. Itier, Using midday surface temperature to estimate daily evaporation from satellite thermal IR data, *Int. J. Remote Sensing*, *4*, 371–383, 1983.

Seguin, B., E. Assad, J. P. Freaud, J. Imbernon, Y. H. Kerr, and J. P. Lagouarde, Use of meteorological satellites for rainfall and evaporation monitoring, *Int. J. Remote Sensing*, *10*, 847–854, 1989.

Seguin, B., J.-P. Lagouarde, and M. Saranc, The assessment of regional crop water conditions from meteorological satellite thermal infrared data, *Remote Sensing Environ.*, *35*, 141–148, 1991.

Sellers, P. J., F. G. Hall, G. Asrar, D. E. Strebel, and R. E. Murphy, The first ISLSCP field experiment (FIFE), *Bull. Am. Meteor. Soc.*, *69*, 22–27, 1988.

Sellers, P. J., S. I. Rasool, and H.-J. Bolle, A review of satellite data algorithms for studies of the land surface, *Bull. Am. Meteor. Soc.*, *71*, 1429–1447, 1990.

Sellers, P. J., M. D. Heiser, and F. G. Hall, Relations between surface conductance and spectral vegetation indices at intermediate ($100\,m^2$ to $15\,km^2$) length scales, *J. Geophys. Res.*, *97*(D17), 19033–19059, 1992.

Sellers, P. J., M. D. Heiser, F. G. Hall, S. J. Goetz, D. E. Strebel, S. B. Verma, R. L. Desjardins, P. M. Schuepp, and J. I. MacPherson, Effects of spatial variability in topography, vegetation cover and soil moisture on area-averaged surface fluxes: A case study using FIFE 1989 data, *J. Geophys. Res.*, *100*(D12), 25607–25629, 1995a.

Sellers, P. J., B. W. Meeson, F. G. Hall, G. Asrar, R. E. Murphy, R. A. Schiffer, F. P. Bretherton, R. E. Dickinson, R. G. Ellingson, C. B. Field, K. F. Huemmrich, C. O. Justice, J. M. Melack, N. T. Roulet, D. S. Schimel, and P. D. Try, Remote sensing of the land surface for studies of global change: Models–algorithms–experiments, *Remote Sensing Environ.*, *51*, 1–17, 1995b.

Shuttleworth, W. J., Aggregation algorithms, *Q. J. R. Meteor. Soc.*, 1998.

Smith, R. C. G., and B. J. Choudhury, Analysis of normalized difference and surface temperature observations over southeastern Australia, *Int. J. Remote Sensing*, *12*, 2021–2044, 1991.

Soer, G. J. R., Estimation of regional evapotranspiration and soil moisture conditions using remotely sensed crop surface temperatures, *Remote Sensing Environ.*, *9*, 27–45, 1980.

Spittlehouse, D. L., and T. A. Black, Evaluation of the Bowen ratio/energy balance method for determining forest evapotranspiration, *Atmos.-Ocean*, *18*, 98–116, 1980.

Stewart, J. B., W. P. Kustas, K. S. Humes, W. D. Nichols, M. S. Moran, and H. A. R. de Bruin, Sensible heat flux-radiometric surface temperature relationship for eight semiarid areas, *J. Appl. Meteor.*, *33*, 1110–1117, 1994.

Sugita, M., and W. Brutsaert, Regional surface fluxes from remotely sensed skin temperature and lower boundary layer measurements, *Water Resour. Res.*, *26*, 2937–2944, 1990.

Sugita, M., W. Brutsaert, Daily evaporation over a region from lower boundary layer profiles, *Water Resour. Res.*, *27*, 747–752, 1991.

Sun, J., and L. Mahrt, Determination of surface fluxes from the surface radiative temperature, *J. Atmos. Sci.*, *52*, 1096–1106, 1995.

Swinback, W. C., The measurement of vertical transfer of heat and water vapour by eddies in the lower atmosphere, *J. Meteor.*, *8*, 135–145, 1951.

Taconet, O., and D. Vidal-Madjar, Applications of a flux algorithm to a field-satellite campaign over vegetated area, *Remote Sensing Environ.*, *26*, 227–239, 1988.

Taconet, O., T. Carlson, R. Bernard, and D. Vidal-Madjar, Evaluation of a surface/vegetation parameterization using satellite measurements of surface temperature, *J. Clim. Appl. Meteor.*, *25*, 1752–1767, 1986.

Troufleau, D., A. Vidal, A. Beaudoin, M. S. Moran, M. A. Weltz, D. C. Goodrich, J. Washburne, and A. F. Rahman, Using optical-microwave synergy for estimating surface energy fluxes over semi-arid rangeland, in *Proceedings of Physical Measurements and Signatures in Remote Sensing*, 17–21 January 1994, Intl. Soc. of Photogrammetry and Remote Sensing (ISPRS), Val d'Isere France, 1994, pp. 1167–1174.

van Bavel, C. H. M., and L. E. Myers, An automatic weighing lysimeter, *Agric. Eng.*, *43*, 580–583, 586–588, 1962.

Wetzel, P. J., D. Atlas, and R. Woodward, Determining soil moisture from geosynchronous satellite infrared data: A feasibility study, *J. Clim. Appl. Meteor.*, *23*, 375–391, 1984.

Wiegand, C. L., A. J. Richardson, D. E. Escobar, and A. H. Gerbermann, Vegetation indices in crop assessments, *Remote Sensing Environ.*, *35*, 105–119, 1991.

Zhan, X., W. P. Kustas, and K. S. Humes, An intercomparison study on models of sensible heat flux over partial canopy surfaces with remotely sensed surface temperature, *Remote Sensing Environ.*, *58*, 242–256, 1996.

Zhang, L., and R. Lemeur, Evaluation of daily evapotranspiration estimates from instantaneous measurements, *Agric. Forest Meteor.*, *74*, 139–154, 1995.

CHAPTER 27

INFILTRATION AND SOIL MOISTURE PROCESSES

PAUL R. HOUSER

Infiltration is the process of water entry from surface sources such as rainfall, snowmelt, or irrigation into the soil. The infiltration process is a component in the overall unsaturated *redistribution* process (Fig. 1)[1] that results in *soil moisture* availability for use by vegetation transpiration, exfiltration (or evaporation) processes, chemical transport, and groundwater recharge. Soil moisture, in turn, controls the partitioning of subsequent precipitation into infiltration and runoff, and the partitioning of available energy between sensible and latent heat flux.

Because of the importance of soil moisture on multiple processes, its definition can be elusive[2]; however, it is most often described as moisture in the unsaturated surface layers (first 1 to 2 m) of soil that can interact with the atmosphere through evapotranspiration and precipitation.[3]

1 CONTROLS ON INFILTRATION AND SOIL MOISTURE

To characterize soil moisture and infiltration, the physical controls on these processes must be considered. The primary soil controls will be considered in this chapter; however, other factors such as soil chemistry, thickness, soil layering or horizons, and preferential flow paths, as well as vegetation cover, tillage, roughness, topography, temperature, and rainfall intensity also exert important controls.[4]

A soil's particle size distribution has a large impact on its hydraulic properties. Soil particles less than 2 mm in diameter are divided into three texture groups (sand, silt, and clay) that help to classify broad soil types and soil water responses (Fig. 2).[5] The type of clay and the coarse material over 2 mm in diameter can also have a

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

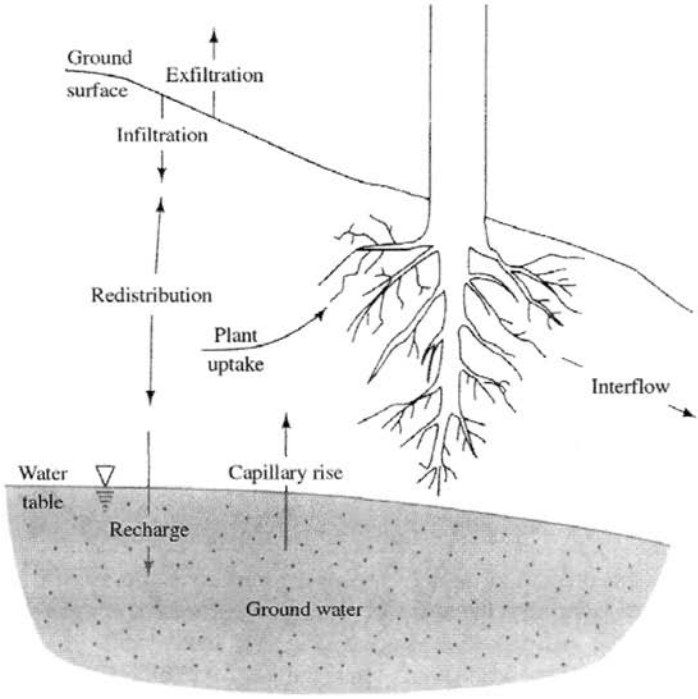


Figure 1 Unsaturated zone definition and active processes.[1]

significant impact on soil water properties. An overview of methods for determining particle size properties is given by Gee and Bauder.[6]

Bulk density, ρ_b (M/L^3) is the ratio of the weight of dry solids to the bulk volume of the soil, and *porosity*, φ (M^3/M^3), is the total volume occupied by pores per unit volume of soil:

$$\varphi = \frac{V_a + V_w}{V_s} = 1 - \frac{\rho_b}{\rho_m} \tag{1}$$

where V_s (L^3) is the total volume of soil, V_a (L^3) is the volume of air, V_w (L^3) is the volume of water, and ρ_m ($M\,L^{-3}$) is the particle density (normally about $2.65\,g/cm^3$).

The volumetric water content, or soil moisture, θ (L^3L^{-3}) is the ratio of water volume to soil volume:

$$\theta = \frac{V_w}{V_s} = \frac{W_w}{W_d}\frac{\rho_b}{\rho_w} \tag{2}$$

where W_w (M) is the weight of water, W_d (M) is the weight of dry soil, and ρ_w (M/L^3) is the density of water. Soil moisture can vary in both time and space, with a

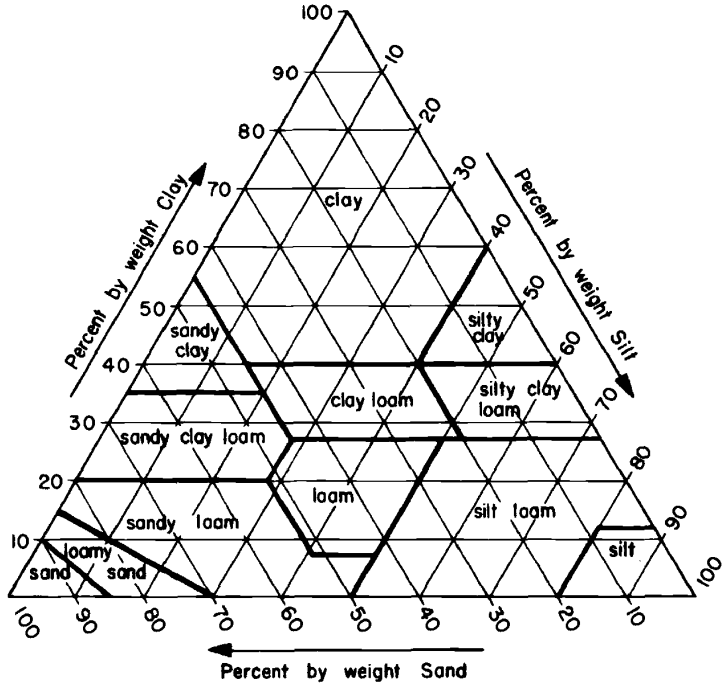


Figure 2 Soil textural triangle describing the relationship between texture and particle size distribution.[5]

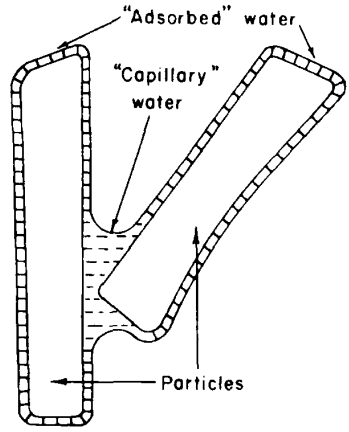


Figure 3 Capillarity and adsorption combine to produce suction.[7]

theoretical range from 0 to φ, but for natural soils the range is significantly reduced due to isolated pore space and tightly held or "adsorbed" water (Fig. 3).[7] If a soil is saturated, then allowed to drain until the remaining water held by surface tension is in equilibrium with gravitational forces, it is at *field capacity*, θ_f. Vegetation can remove water from the soil until the *permanent wilting point*, θ_w, is reached. Therefore, the *available water content* for plant use, $\theta_a = \theta_f - \theta_w$. Typical ranges of porosity, field capacity, and wilting point for different soils are given in Fig. 4.[8]

In unsaturated soils, water is held in the soil against gravity by surface tension (Fig. 3). This tension, suction, or *matric potential*, ψ (L), increases as the radii of curvature of the meniscus or water content decreases (Fig. 5).[9] Matric potential is expressed in reference to atmospheric pressure, so for saturated soil $\psi = 0$ and for unsaturated soil $\psi < 0$.

The *hydraulic conductivity*, K (L/T), is a measure of the ability of the soil to transmit water that varies nonlinearly over a large range depending on both soil

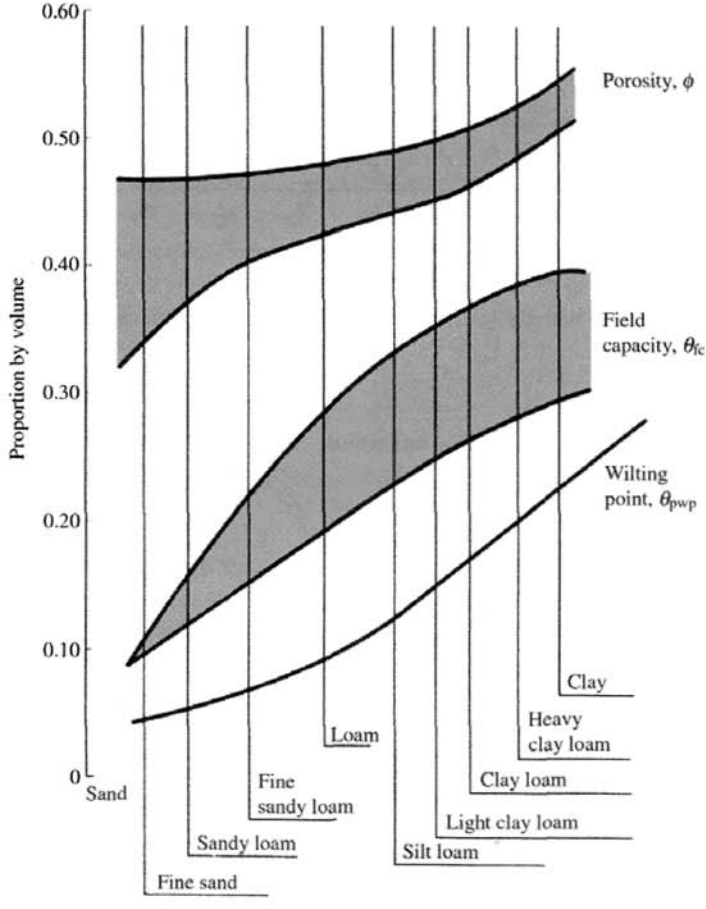


Figure 4 Water holding properties of various soils.[8]

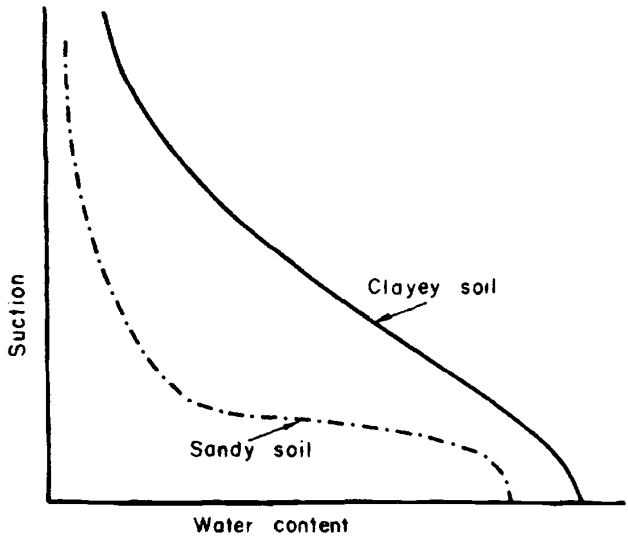


Figure 5 Effect of texture on water retention characteristics.[9]

properties and water content (Fig. 6).[10] Many laboratory and field hydraulic conductivity measurement methods exist for use with various soils; see Bouwer and Jackson[11] or Green et al.[12] for details.

Soil water content can significantly impact infiltration by (1) increasing the hydraulic conductivity, which increases infiltration, and (2) reducing the surface tension that draws moisture into the soil, which reduces infiltration. The net effect of these impacts depends on the water content itself, the water input rate, and duration and the distribution of hydraulic conductivity.

The *water retention characteristic* describes a soil's ability to store and release water and is defined by the relationship between soil moisture and the matric potential (Fig. 5). This is a power function relationship that has been described by Brooks and Corey[13] and Van Genuchten,[14] among others. The water tension characteristic is usually measured in air pressure chambers where the water content of a soil sample can be monitored over a wide pressure range.[15]

The water retention relationship may actually change between drying and wetting due to the entrapment of air in soil pores (Fig. 7).[16] For practical applications, this effect, called *hysteresis*, is usually neglected.[17]

2 PRINCIPLES OF SOIL WATER MOVEMENT

Through experiments on saturated water flow through sand beds, Darcy[18] found that the rate of flow, Q (L^3/T), through a cross-sectional area A (L^2), is directly propor-

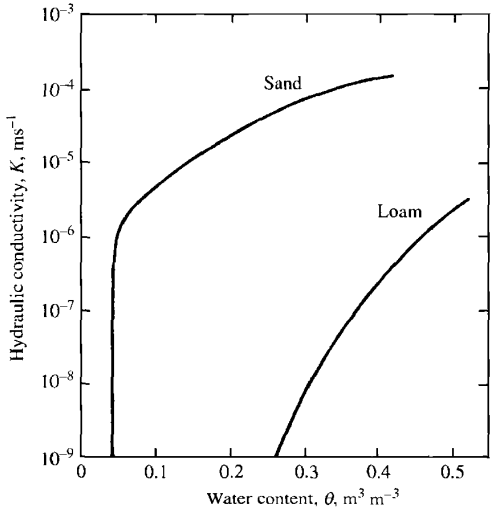


Figure 6 Effect of texture and soil moisture on hydraulic conductivity.[10]

tional to head loss (e.g., water elevation difference), ΔH (L), and inversely to the flow path length, Δl (L):

$$Q = KA\frac{\Delta H}{\Delta l} \tag{3}$$

Combining *Darcy's law* with the law of conservation of mass results in a description of unsaturated flow called *Richards equation*[19]:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left(\frac{K}{C}\frac{\partial \theta}{\partial z}\right) - \frac{\partial K}{\partial z} \tag{4}$$

where $C = -\partial\theta/\partial\psi$ is the water content change in a unit soil volume per unit matric potential, ψ change. The Richards equation is the basis for most simulations of infiltration and redistribution of water in unsaturated soil. Using some approximations, analytical solutions of the Richards equation are available[20,21] that show good agreement with observations.[22] The Richards equation is based on saturated flow theory, and does not account for all of the processes active in natural systems, so it may not always perform well.[23]

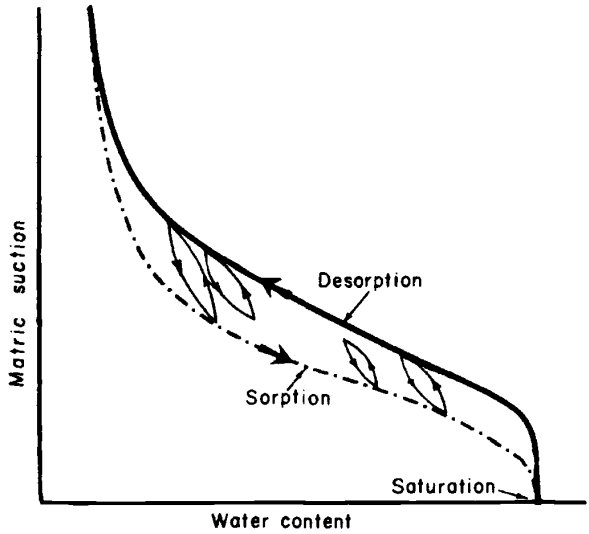


Figure 7 Changes in water retention characteristics between sorption and desorption.[16]

3 INFILTRATION ESTIMATION

Some basic principles that govern the movement of water into the soil can be used to predict infiltration. The *infiltration capacity*, f (L), is the maximum rate that a soil in a given condition can absorb water and generally decreases as soil moisture increases. If the *rainfall rate* is less than the infiltration capacity, then infiltration proceeds at the capacity rate. However, if the rainfall rate exceeds the infiltration capacity, then infiltration proceeds at the capacity rate, and the excess rainfall ponds on the surface or runs off. As the time from the onset of rainfall increases, infiltration rates decrease due to soil moisture increases, raindrop impact, and the clogging of soil pores, until a steady-state infiltration rate is reached (Fig. 8).[24] Existing infiltration models use empirical, approximate, or physical approaches to predict infiltration.[25]

Empirical. Empirical infiltration models generally utilize a mathematical function whose shape as a function of time, t, matches observations and then attempts a physical explanation of the process.

Kostiakov[26] proposed the simple infiltration rate, f (L/T) model:

$$f = \alpha\gamma t^{\alpha-1} \tag{5}$$

where α and γ are constants that have no particular meaning and must be evaluated by fitting the model to experimental data.

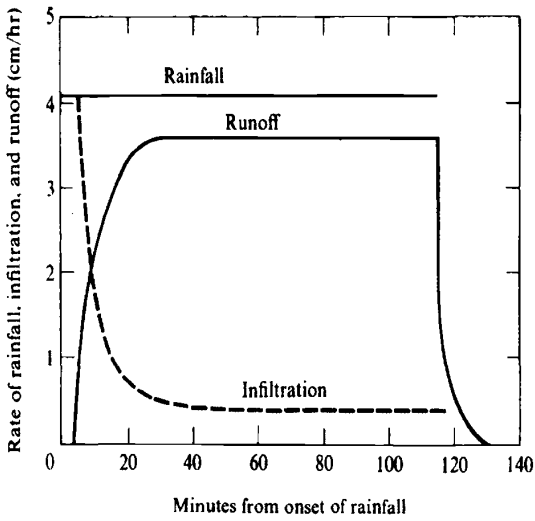


Figure 8 Idealized relationship between rainfall, infiltration, and runoff rates.[24]

Horton's[27] infiltration model has been widely used in hydrologic simulation. It relates infiltration capacity to initial infiltration rate, and f_0, the constant infiltration rate at large times, f_c:

$$f = f_c + (f_0 - f_c)e^{-\beta t} \tag{6}$$

where β is a soil parameter describing the rate of decrease of infiltration.

Approximate. Analysis approximations to the Richards equation are possible if several simplifying assumptions are made. Most approximate infiltration models treat the soil as a semi-infinite medium, with the soil saturating above a wetting front.

Green and Ampt[28] assumed in a soil with constant hydraulic properties, the matric potential at the moving wetting front is constant, leading to a discontinuous change in soil moisture at the wetting front:

$$f = K\left[1 + \frac{(\varphi - \theta_i)S_f}{F}\right] \tag{7}$$

where S_f (L) is the effective suction at the wetting front, θ_i is the initial water content, and F (L) is the accumulated infiltration.

Phillip[29] proposed that the first two terms in a series of powers of $t^{1/2}$ could be used to approximate infiltration:

$$f = \tfrac{1}{2}St^{1/2} + A \tag{8}$$

where S is a parameter called sorptivity, t is time from ponding, and A is a constant that depends on soil properties. In this model, the infiltration rate approaches a constant equal to the hydraulic conductivity at the surface water content, and the wetting front advances without changing its shape and approaches a constant velocity.

Physical. Recent advances in numerical methods and computing has facilitated the practical application of the Richards equation to realistic flow problems. Such packages can simulate water infiltration and redistribution using the Richards equation and including precipitation, runoff, drainage, evaporation, and transpiration processes.[30]

4 INFILTRATION MEASUREMENT

Infiltration rates can be measured at a point using a variety of methods described here, each appropriate for certain conditions. However, because of the large temporal and spatial variability of infiltration processes, catchment average infiltration rates may be desired, which can be obtained through the water balance analysis of rainfall–runoff observations.[31]

Ring Infiltrometer. This simple method is most appropriate for flood irrigation or pond seepage infiltration. A cylindrical metal ring is sealed at the surface and flooded. Intake measurements are recorded until steady-state conditions are reached.[32] If the effects of lateral flow are significant, then a double-ring infiltrometer can be used. Due to ponding conditions within the ring, observed infiltration rates are often higher than under natural conditions.[33]

Sprinkler Infiltrometer. This method is appropriate for quantifying infiltration from rainfall. Artificial rainfall simulators are used to deliver a specified rainfall rate to a well-defined plot. Runoff from the plot is measured, allowing computation of the infiltration rate.[34,35]

Tension Infiltrometer. The tension or disk infiltrometer employs a soil contact plate and a water column that is used to control the matric potential of the infiltrating water. By varying the tension, the effect of different size macropores can be determined.[36,37]

Furrow Infiltrometer. This method is useful if information on infiltration of flowing water in irrigation furrows is desired. Either the water added to a small section of blocked off furrow to maintain a constant depth or the inflow–outflow of a furrow segment can be monitored to determine the infiltration characteristics of the system.[38]

5 SOIL MOISTURE MEASUREMENT

Soil water content can be determined directly using gravimetric techniques or indirectly by inferring it from a property of the soil.[39,40]

Gravimetric. The oven-drying soil moisture measurement technique is the standard for calibration of all other methods but is time consuming and destructive. The method involves obtaining a wet soil sample weight, drying the sample at 105°C for 24 h, then obtaining the dry sample weight [see Eq. (2)].

Neutron Thermalization. High-energy neutrons are emitted by a radioactive source into the soil and are preferentially slowed by hydrogen atoms. The number of slow neutrons returning to the detector are a measure of soil moisture.

Gamma Attenuation. The attenuation in soil of gamma rays emitted from caesium-137 is directly related to soil density. If the soil's bulk density is assumed to be constant, then changes in attenuation reflect changes in soil moisture.[41]

Time-Domain Reflectometry (TDR). TDR measures the soil's dielectric constant, which is directly related to soil moisture, by measuring the transmit time of a voltage pulse applied to a soil probe.

Tensiometric Techniques. This method measures the capillary or moisture potential through a liquid-filled porous cup connected to a vacuum gage. Conversion to soil moisture requires knowledge of the water retention characteristic.

Resistance. The electrical resistance or conductivity of a porous block (nylon, fiberglass, or gypsum) imbedded in the soil depends primarily on the water content of the block. However, because of salinity and temperature sensitivity, measurements of these sensors are of limited accuracy.[42]

Heat Dissipation. Changes in the thermal conductivity of a porous block imbedded in the soil depend primarily on the water content of the block. The dissipation of a heat pulse applied to the block can be monitored using thermistors, then the soil water content can be determined from calibration information.

Remote Sensing. Soil moisture can be remotely sensed with just about any frequency where there is little atmospheric absorption.[43] But, it is generally accepted that long wavelength, passive microwave sensors have the best chance of obtaining soil moisture measurements that contain little error introduced by vegetation and roughness and offer great potential to remotely sense soil moisture content with depth due to differential microwave absorption with varying dielectric constant.[44]

6 SPATIAL AND TEMPORAL VARIABILITY

Natural soils exhibit considerable spatial heterogeneity in both the horizontal and vertical directions, and at all distance scales from the pore to the continent, to a degree that it is difficult to capture this variability in routine measurements.[45,46] This large variation in soil properties, infiltration, and soil moisture over relatively small areas makes it difficult to transfer the understanding of processes developed at a point to catchment scales. Many hydrological models assume that a single spatially representative average soil property can be used to characterize catchment (or even larger) scale processes. It is clear from the nonlinear character of soil water processes [Eq. 94)] that catchment average infiltration cannot be computed based on catchment average soil properties. It is also clear that the physical meaning of a soil property, say porosity, is relative to the volume over which it is averaged.[47] However, there is a need to understand and reduce this complexity for the purposes of prediction and management. Several approaches, including dividing the catchment into hydrologically similar subareas,[48] various statistical approaches,[49] and scaling and similarity theory[50,51] have made headway toward an understanding of infiltration and soil moisture spatial variability, but are not being widely used in practical applications.

One of the most important recent findings in this regard is the scale invariance of soil water behavior. If a heterogenous field is the union of homogenous spatial domains, each with associated characteristic length scales, then heterogeneity simplifies into the spatial variability of these length scales, while the functional relationships that describe soil water movement (i.e., the Richards equation) remain uniform across spatial scales.[52] This new understanding of the underlying symmetry of the Richards equation may help to facilitate a workable scale invariant analytical soil water dynamical model.

Finally, there is a continuing need for the observation of soil properties, soil moisture, and infiltration processes at multiple scales to facilitate understanding and prediction of these complex and socially significant processes. It is likely that remote sensing of soil moisture and other land surface factors will be instrumental in this respect.

REFERENCES

1. Dingman, S. L., *Physical Hydrology*, Prentice-Hall, Englewood Cliffs, NJ, 1994, p. 211.
2. Lawford, R. G., An overview of soil moisture and its role in the climate system, in F. J. Eley, R. Granger, and L. Martin (Eds.), *Soil Moisture; Modeling and Monitoring for Regional Planning*, National Hydrology Research Centre Symposium No. 9 Proceedings, 1992, pp. 1–12.
3. Schmugge, T., T. J. Jackson, and H. L. McKim, Survey of in-situ and remote sensing methods for soil moisture determination, in *Satellite Hydrology*, American Water Resources Association, 1979.
4. Rawls, W. J., L. R. Ahuja, D. L. Brakensiek, and A. Shirmohammadi, Infiltration and soil water movement, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, pp. 5.1–5.51.

5. Hillel, D., *Introduction to Soil Physics*, Academic, New York, 1982, p. 29.
6. Gee, G. W., and J. W. Bauder, Particle size analysis, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 383–411.
7. Hillel, D., *Fundamentals of Soil Physics*, Academic, New York, 1980, p. 143.
8. Dunne, T., and L. Leopold, *Water in Environmental Planning*, W. H. Freeman, New York, 1978, p. 175.
9. Hillel, D., *Fundamentals of Soil Physics*, Academic, New York, 1980, p. 150.
10. Marshall, T. J., and J. W. Holmes, *Soil Physics*, 2nd ed., Cambridge University Press, 1988, p. 87.
11. Bouwer, H., and R. D. Jackson, Determining soil properties, in J. van Schilfgaard (Ed.), *Drainage for Agriculture*, American Society of Agronomy, Madison, WI, 1974, pp. 611–672.
12. Green, R. E., L. R. Ahuja, and S. K. Chong, Hydraulic conductivity, diffusivity, and sorptivity of unsaturated soils—field methods, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 771–798.
13. Brooks, R. H., and A. T. Corey, *Hydraulic Properties of Porous Media*, Hydrology Paper 3, Colorado State University, Fort Collins, CO, 1964.
14. Van Genuchten, M. Th., A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Am. J., 44*, 892–898, 1980.
15. Cassel, D. K., and A. Klute, Water potential: Tensiometry, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 563–596.
16. Hillel, D., *Fundamentals of Soil Physics*, Academic, New York, 1980, p. 153.
17. Rawls, W. J., L. R. Ahuja, D. L. Brakensiek, and A. Shirmohammadi, Infiltration and Soil Water Movement, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, pp. 5.5–5.6.
18. Darcy, H., *Les fontaines publiques de la ville de Dijon*, Dalmont, Paris, 1856.
19. Richards, L. A., Capillary conduction of liquids in porous mediums, *Physics, 1*, 318–333, 1931.
20. Kuhnel, V., V. C. I. Dooge, G. C. Sander, and J. P. J. O'Kane, Duration of atmosphere-controlled and soil-controlled phases of infiltration for constant rainfall at a soil surface, *Ann. Geophys., 8*, 11–20, 1990.
21. Sposito, G., Recent advances associated with soil water in the unsaturated zone, in *Reviews of Geophysics*, Supplement, U.S. National Report to International Union of Geodesy and Geophysics 1991–1994, 1995, pp. 1059–1065.
22. Whisler, F. D., and H. Bouwer, A comparison of methods for calculating vertical drainage and infiltration in soils, *J. Hydrol., 10*, 1–19, 1970.
23. Nielsen, D. R., J. W. Biggar, and J. M. Davidson, Experimental consideration of diffusion analysis in unsaturated flow problems, *Soil Soc. Am. Proc., 26*, 107–111, 1962.
24. Dunne, T., and L. Leopold, *Water in Environmental Planning*, W. H. Freeman, New York, 1978, p. 169.
25. Rawls, W. J., L. R. Ahuja, D. L. Brakensiek, and A. Shirmohammadi, Infiltration and soil water movement, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, pp. 5.21–5.23.

26. Kostiakov, A. N., On the dynamics of the coefficient of water-percolation in soils and on the necessity for studying it from a dynamic point of view for purposes of amelioration, *Trans. Sixth Comm. Intern. Soil. Sci. Soc. Russian*, part A, 1932, pp. 17–21.
27. Horton, R. E., An approach toward a physical interpretation of infiltration-capacity, *Soil Sci. Soc. Am. J.*, *5*, 399–417, 1940.
28. Green, W. H., and G. A. Ampt, Studies on soil physics: 1. Flow of air and water through soils, *J. Agric. Sci.*, *4*, 1–24, 1911.
29. Phillip, J. R., The theory of infiltration: 1. The infiltration equation and its solution, *Soil Sci.*, *83*, 345–357, 1957.
30. Ross, O. J., Efficient numerical methods for infiltration using Richards equation, *Water Resour. Res.*, *26*, 279–290, 1990.
31. Soil Conservation Service, Hydrology, in *SCS National Engineering Handbook*, U.S. Department of Agriculture, Washington, DC, 1972, Sec. 4.
32. Bouwer, H., Intake rate: Cylinder infiltrometer, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 825–844.
33. Lukens, R. P. (Ed.), *Annual Book of ASTM Standards*, Part 19: *Soil and Rock, Building Stones*, 1981, pp. 509–514.
34. Agassi, M., I. Shainberg, and J. Morin, Effects on seal properties of changes on drop energy and water salinity during a continuous rainstorm, *Aust. J. Soil Res.*, *26*, 1–10, 1988.
35. Peterson, A., and G. Bubenzer, Intake rate sprinkler infiltrometer, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 845–870.
36. Ankeny, M. D., T. C. Kaspar, and R. Horton, Design for an automated tension infiltrometer, *Soil Sci. Soc. Am. J.*, *52*, 893–896, 1988.
37. Perrouix, K. M., and I. White, Designs of disc permeameters, *Soil Sci. Soc. Am. J.*, *52*, 1205–1215, 1988.
38. Kincaid, D. C., Intake rate: Border and furrow, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 871–887.
39. Gardner, W. H., Water content, in A. Klute (Ed.), *Methods of Soil Analysis, Part I—Physical and Mineralogical Methods*, 2nd ed., American Society of Agronomy Monograph 9, 1986, pp. 493–544.
40. Schmugge, T. J., T. J. Jackson, and H. L. McKim, Survey of methods for soil moisture determination, *Water Resour. Res.*, *16*(6), 961–979, 1980.
41. Jury, W. A., W. R. Gardner, and W. H. Gardner, *Soil Physics*, 5th ed., Wiley, 1991, pp. 45–47.
42. Hillel, D., *Introduction to Soil Physics*, Academic, New York, 1982, pp. 57–89.
43. Schmugge, T. J., Remote sensing of soil moisture, in *Hydrological Forecasting*, Wiley, New York, 1985.
44. Ulaby, F. T., G. A. Bradley, and M. C. Dobson, Potential application of satellite radar to monitor soil moisture, in *Satellite Hydrology*, American Water Resources Association, 1979, pp. 363–370.
45. Jury, W. A., W. R. Gardner, and W. H. Gardner, *Soil Physics*, 5th ed. Wiley, New York, 1991, pp. 268–293.

46. Dingman, S. L., *Physical Hydrology*, Prentice-Hall, Englewood Cliffs, NJ, 1994, pp. 247–250.
47. Bear, J., *Dynamics of Fluids in Porous Media*, Elsevier, New York, 1972.
48. Springer, E. P., and T. W. Cundy, Field-scale evaluation of infiltration parameters from soil texture for hydrologic analysis, *Water Resour. Res.*, *23*, 325–334, 1987.
49. Berndtsson, R., and M. Larson, Spatial variability of infiltration in a semi-arid environment, *J. Hydrol.*, *90*, 117–133, 1987.
50. Sharma, M. L., G. A. Gander, and G. C. Hunt, Spatial variability of infiltration in a watershed, *J. Hydrol.*, *45*, 101–122, 1980.
51. Wood, E. F., M. Sivaplan, and K. Beven, Similarity and scale in catchment storm response, *Rev. Geophys.*, *28*, 1–18, 1990.
52. Sposito, G., Recent advances associated with soil water in the unsaturated zone, in *Reviews of Geophysics*, Supplement, U.S. National Report to International Union of Geodesy and Geophysics 1991–1994, 1995, pp. 1059–1065.

CHAPTER 28

GROUNDWATER FLOW PROCESSES

WILLIAM W-G. YEH

1 INTRODUCTION

In the hydrologic cycle, groundwater occurs whenever surface water occupies and saturates the pores or interstices of the rocks and soils beneath Earth's surface. The geologic formations that store and transmit the subsurface water are known as aquifers. Aquifers, aquitards (semipermeable formations), or aquicludes (nonpermeable formations) may underlie a geographic area, watershed, or drainage basin, and all may hold water. But drawing water from aquitards and aquicludes is impractical and economically prohibitive, whereas groundwater stored in aquifers can be removed economically and is often a dependable source of water supply (Todd, 1980). Most aquifers can be considered as underground storage reservoirs that receive recharge from both natural and artificial sources.

Depending on local geological formation and boundary conditions, groundwater may flow out of the aquifer, contributing to surface runoff. In most cases, each aquifer formation has spatially varying properties, such as transmissivity and storativity, which affect the basin's response to pumping and artificial recharge. These formations are collectively referred to as a groundwater reservoir or groundwater system (Willis and Yeh, 1987). Groundwater aquifers can be classified as confined or unconfined, depending on the existence of a water table. A leaky confined aquifer represents a geological formation that leaks and allows water to flow through the confining layer.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

2 DARCY'S LAW

The fundamental law that governs groundwater flow in a laminar flow regime is Darcy's law. If we assume the porous medium is homogeneous and isotropic, Darcy's law states that the specific discharge is proportional to the gradient of hydraulic head:

$$\mathbf{q} = -K\nabla h \tag{1}$$

where $\mathbf{q}$ is the specific discharge vector (volume flow rate per unit cross-sectional area normal to the direction of flow), K is the hydraulic conductivity, h is the head, and ∇h is the gradient vector of the head,

$$\nabla h = \left(\frac{\partial h}{\partial x}i + \frac{\partial h}{\partial y}j + \frac{\partial h}{\partial z}k\right) \tag{2}$$

where i, j, and k are unit vectors in the x, y, and z coordinate directions, respectively. The hydraulic conductivity (K) is a function of both fluid and medium properties. As can be shown by dimensional analysis using the basic units of length (L), mass (M), and time (T), K can be expressed as (see, e.g., DeWiest, 1965):

$$K = \frac{Cd^2\gamma}{\mu} = k\frac{\gamma}{\mu} \tag{3}$$

where d (L) is some characteristic length of the medium, e.g., the average pore size or mean grain diameter of the granular material, μ M/(LT) is the dynamic viscosity, γ M/(L^{2}T^2) is the specific weight of the fluid (water), and C is a constant or shape factor, which accounts for the effects of stratification, packing, arrangement of grains, size distribution, and porosity. Parameter k is referred to as the intrinsic permeability and is solely dependent on the medium properties ($k = Cd^2$).

The porous medium is said to be homogeneous if the hydraulic conductivity is independent of the position (x, y, z) within the aquifer. If not, the aquifer is inhomogeneous, i.e., $K = K(x, y, z)$. The isotropy or anisotropy of the aquifer reflects the directional variability of the hydraulic conductivity. If the hydraulic conductivity varies with the direction of flow, the aquifer is anisotropic. On the other hand, if the hydraulic conductivity is independent of the direction of flow, the aquifer is isotropic. The conditions of inhomogeneity and anisotropy are common occurrences in the soils and geologic formations of aquifers.

Because the specific discharge may not be collinear with the gradient of the hydraulic head, nor have equal specific discharge components in the x, y, and z

directions, the hydraulic conductivity may be represented as a second-order tensor quantity (Eagleson, 1970). Darcy's law can then be generalized as:

$$\begin{bmatrix} q_x \\ q_y \\ q_z \end{bmatrix} = -\begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix} \begin{bmatrix} \dfrac{\partial h}{\partial x} \\ \dfrac{\partial h}{\partial y} \\ \dfrac{\partial h}{\partial z} \end{bmatrix} \tag{4}$$

If the coordinate system is aligned with the principal directions of the hydraulic conductivity tensor, Darcy's law in three dimensions can be written as:

$$q_x = -K_{xx}\frac{\partial h}{\partial x} \tag{5}$$

$$q_y = -K_{yy}\frac{\partial h}{\partial y} \tag{6}$$

$$q_z = -K_{zz}\frac{\partial h}{\partial z} \tag{7}$$

The governing equations for groundwater flow are generally derived by combining Darcy's law with the continuity equation (conservation of mass).

3 FLOW EQUATION FOR A CONFINED OR LEAKY AQUIFER

In a confined aquifer, the amount of water released from groundwater storage is dependent on the compressibility of the water and of the porous medium. Confined aquifers are bounded above and below by confining layers. In contrast, leaky or semiconfined aquifers have semipermeable confining layers that are capable of leakage and storage. A multilayered aquifer system is a system in which the aquifers are hydraulically interdependent as changes in head in one layer, caused by pumping, or recharge, can induce flow to and from adjacent layers. If we assume that the leakage or flow between layers and aquitards occurs only in the vertical direction and that any storage effects in the aquitards are negligible, the governing equation characterizing two-dimensional horizontal flow in a semiconfined or leaky aquifer can be expressed by

$$\frac{\partial}{\partial x}\left(T_{xx}\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(T_{yy}\frac{\partial h}{\partial y}\right) = S\frac{\partial h}{\partial t} - L \pm Q \tag{8}$$

where T_{xx}, T_{yy} = components of transmissivity along the x and y coordinate axes, the product of the hydraulic conductivity and the aquifer's thickness (L^2/T)
S = storage coefficient (dimensionless)
h = head in the main aquifer (L)
t = time (L)
L = leakage from the overlying semiconfining stratum (L/T)
Q = sink/source term (L/T)

The leakage term can be calculated by Darcy's law:

$$L = K_z \frac{H - h}{b'} \tag{9}$$

where K_z = vertical hydraulic conductivity of the overlaying semiconfining stratum (L/T)
b' = thickness of the overlying semiconfining stratum (L)
H = external head in the overlying semiconfining stratum (L)

The effect of pumping and injection wells in the main aquifer can be simulated by representing the wells as point sources or point sinks under the assumption that the wells fully penetrate the thickness of the aquifer. If we let the index set Ω be the locations of all the pumping and injection wells, then the point sources/sinks can be expressed as:

$$\pm \sum_{w \in \Omega} Q_w \delta(x - x_w, y - y_w) \tag{10}$$

where $+Q_w$ is the pumping ($-Q_w$ recharge) from the wth pumping (injection) well located at (x_w, y_w) and $\delta(x - x_w, y - y_w)$ is the Dirac delta function, where

$$\delta(x - x_w, y - y_w) = \begin{cases} 1, & \text{if } x = x_w, y = y_w \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Equation (8) applies to a leaky aquifer system in which the main aquifer is overlain by a semiconfining stratum that leaks water into the main aquifer through the semiconfining stratum. If the term L goes to zero, Eq. (8) applies to a strictly confined system.

4 FLOW EQUATION FOR AN UNCONFINED AQUIFER

In contrast to confined aquifers, an unconfined aquifer has a free surface (water table) boundary, a boundary at atmospheric pressure. Water released from storage

occurs due to gravity drainage as the water table in the aquifer responds to pumping, drainage, or natural or artificial recharge. The unconfined flow problem is commonly analyzed using the Dupuit assumptions: (1) uniform and horizontal flow within any vertical cross section, and (2) the velocity at the free surface may be expressed as $q_x = -K(\partial h/\partial x)$. The second assumption implies small slopes of the free surface.

Using the concept of vertical averaging, the governing equation characterizing two dimensional horizontal flow in an unconfined aquifer can be expressed as:

$$\frac{\partial}{\partial x}\left(K_{xx}h\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}h\frac{\partial h}{\partial y}\right)=[S_y+S_sh]\frac{\partial h}{\partial t}-R \tag{12}$$

where K_{xx}, K_{yy} = components of hydraulic conductivity along the x and y coordinate axes (L/T)
S_y = specific yield of the aquifer (dimensionless)
S_s = specific storage of the aquifer (1/L)
R = net recharge (L/T)

The specific storage effect is generally negligible when compared to the specific yield and can be dropped to give

$$\frac{\partial}{\partial x}\left(K_{xx}h\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}h\frac{\partial h}{\partial y}\right)=S_y\frac{\partial h}{\partial t}-R \tag{13}$$

which is the nonlinear Boussinesq equation. Pumping and injection wells may also be incorporated via Eq. (10) in the recharge term of the equation as point sources or sinks. There are several ways to linearize Eq. (13). The first method is based on the assumption that the depth of the flow varies slightly in the flow domain, e.g., mildly sloping aquifers. The head may then be expressed by

$$h=\bar{h}+\hat{h} \tag{14}$$

where $\bar{h}$ is the average depth of flow and $\hat{h}$ is the derivation of the head from $\bar{h}$. If we assume $\hat{h} \ll \bar{h}$, the Boussinesq equation becomes

$$\frac{\partial}{\partial x}\left(K_{xx}\bar{h}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\bar{h}\frac{\partial h}{\partial y}\right)=S_y\frac{\partial h}{\partial t}-R \tag{15}$$

or

$$\frac{\partial}{\partial x}\left(T_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(T_{yy}\frac{\partial h}{\partial y}\right)=S_y\frac{\partial h}{\partial t}-R \tag{16}$$

where $T_{xx}=K_{xx}\bar{h}$ and $T_{yy}=K_{yy}\bar{h}$. It can be seen that Eq. (16) is identical to the governing equation of the confined flow.

The second method of linearization is based on the temporal variation of the temporal derivative. Rewriting the temporal derivative as

$$S_y \frac{\partial h}{\partial t} = \frac{S_y}{2h} \frac{\partial h^2}{\partial t} \tag{17}$$

and assuming $\bar{S} = S_y/2h$ is approximately constant and equal to $S_y/2\bar{h}$, the Boussinesq equation is intrinsically linear in h^2,

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h^2}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h^2}{\partial y}\right) = 2\bar{S}\frac{\partial h^2}{\partial t} - 2R \tag{18}$$

If the initial and boundary conditions are also linear in h^2, Eq. (18) has been shown to be more accurate in predicting the water level.

The third method of linearization, as is used in MODFLOW (McDonald and Harbaugh, 1988), is to use the calculated head value from the last iteration h' to replace h, i.e., $T_{xx} = K_{xx}h'$ and $T_{yy} = K_{yy}h'$, and iteration continues until a convergence criterion is met.

5 INITIAL AND BOUNDARY CONDITIONS

The most common types of boundary conditions are Dirichlet, Neumann, and Cauchy.

1. Dirichlet conditions occur when a portion of the boundary is at a prescribed head level. For example, if an aquifer is adjacent to a stream or lake, then

$$h(x, y, t) = f_1(x, y, t) \qquad (x, y) \in \partial\Omega_1 \tag{19}$$

   where f_1 is a known function and $\partial\Omega_1$ is the boundary.
2. Neumann conditions occur when a portion of the boundary has a specified flow transversing it normal to the boundary. For example, if a portion of the aquifer boundary is subject to recharge, then

$$T\frac{\partial h}{\partial n} = f_2(x, y, t) \qquad (x, y) \in \partial\Omega_2 \tag{20}$$

   where f_2 is a known function and $\partial h/\partial n$ is the normal derivative to the boundary $\partial\Omega_2$.
3. The Cauchy, or mixed boundary, condition occurs when both the head and its normal derivative are specified on the boundary $\partial\Omega_3$, for example, the induced infiltration in a coupled aquifer and stream system (Prickett and Lonnquist, 1971).

Prior to solving the groundwater flow equation, the initial condition must be specified in the flow domain.

Typically, the initial condition can be expressed by

$$h(x, y, 0) = f_3(x, y) \qquad (x, y) \in \Omega \tag{21}$$

where Ω is the flow domain and f_3 is a known function.

6 DATA COLLECTION

Before a groundwater model can be used for prediction or management purposes, it must be calibrated using historical observations, historical operational records, and initial and boundary conditions. Calibration, in the general sense, concerns the estimation of model parameters (parameter identification) in a given conceptual model. Without doubt, the more data (observations) we have, the more reliable the calibrated model will be. Typical data available include the following:

1. *Topographical and geographical maps*. This information assists in defining the region of the groundwater basin and boundary types, i.e., impermeable, given head, given flux, and so on.
2. *Well logs*, which contain the vertical distribution of geological formations, including depth, color, character, size of material, and structure of the strata. This information helps to determine the layer structure and parameter value range in each layer.
3. *Groundwater level observations*. This information is most crucial to model calibration, as most of the groundwater models are calibrated against historical water level observations.
4. *Historical precipitation and streamflow data*. This information provides recharge and boundary conditions.
5. *Historical groundwater pumping/injection records*. This information provides the sink/source term.
6. *Land use data*, which contain acreages of human activities, such as residential, commercial, industrial, and agricultural uses. This information helps to estimate the groundwater consumption and return flow.

7 SELECTION OF NUMERICAL MODELS

Analytical solutions of groundwater flow have been reported for idealized groundwater systems; however, a more common approach to solving the distributed-parameter, time-dependent partial differential equations that govern the groundwater flow is through numerical techniques such as the finite-difference or finite-element methods. These techniques transform the partial differential equations into a system

of algebraic equations. The solution of the system of algebraic equations determines the head values at a predetermined set of discrete nodal points within the aquifer system.

Finite-difference approximation is based on the Taylor series representation of the time and spatial derivatives. It is conceptually more straightforward than the finite-element approximation and easy to implement. Finite-element approximation is based on the method of weighted residuals. For many groundwater problems, the finite-element method may be more advantageous over the finite difference-method. Medium heterogeneity and irregular boundary conditions are handled easily by the finite element method. This contrasts with finite-difference approximation that requires complicated interpolation schemes to approximate complex boundary conditions. Moreover, in the finite-element method, the size of the elements can easily be varied to reflect rapidly changing state variables or parameter values. The piecewise continuous representation of the dependent variables and, possibly, the parameters of the groundwater system can also increase the accuracy of numerical approximation (Willis and Yeh, 1987).

Since these two types of numerical methods have been applied to many fields, references are abundant in the literature. Willis and Yeh (1987), Anderson and Woessner (1992), and Sun (1994b) have provided detailed analyses of how these two methods are to be applied to groundwater modeling. It is also worth noting that many established groundwater modeling softwares are available in the public domain. Bedient et al. (1994) provided a summary listing of existing numerical models of groundwater flow and solute transport.

8 PARAMETER ESTIMATION (PARAMETER IDENTIFICATION)

Thc accuracy of model prediction depends on the reliability of the estimated model parameters as well as on the accuracy of the prescribed initial and boundary conditions. In general, parameters used in deriving the governing equations are not directly measurable from the physical point of view. In practice, model parameters are required to be estimated from historical input–output observations using an inverse procedure of parameter estimation.

The inverse problem of parameter estimation in distributed-parameter systems has been studied extensively during the last three decades. The term distributed system implies that the response of the system is governed by a partial differential equation [(Eq. (8) or (13)] and parameters embedded in the equation (T_{xx}, T_{yy}, S) are spatially dependent. A review of the inverse problem of parameter identification in groundwater hydrology was presented by Yeh (1986), Carrera (1988), Sun (1994a), and McLaughlin and Townley (1996). In general, the inverse problem seeks to identify the model parameters by observing the output of the dependent variable (head) in the spatial and time domain. Frequently, point estimates of transmissivity and storage coefficient are also available and they can be used as prior information to regulate the inverse solution.

9 PARAMETERIZATION

The number of observations is finite and limited, whereas the spatial domain is continuous. For an inhomogeneous aquifer, the dimension of, for example, the transmissivity is theoretically infinite. In practice, the infinite parameter dimension must be reduced to a finite dimensional form. The reduction of the number of parameters from the infinite dimension to a finite dimensional form is called parameterization (Yeh and Yoon, 1976, 1981; Yeh, 1986; Sun, 1994a). Parameterization can be achieved by either a deterministic method or by a stochastic model. In general, parameterization can be achieved by the following methods.

Zonation Method

In this approach, the flow region of the aquifer is divided into a number of zones, and a constant parameter value is used to characterize the aquifer property in each zone. The unknown transmissivity function is then represented by a number of constants, which is equal to the number of zones. Here, we mention the work of Coats et al. (1970), Emsellem and de Marsily (1971), Yeh and Yoon (1976), and Cooley (1977, 1979).

In principle, the zonation pattern and its corresponding parameter values should be determined simultaneously (Sun and Yeh, 1985; Sun et al., 1998).

Interpolation Method

If, for example, finite elements are used as the interpolation method, the unknown parameter distribution in the flow region is discretized into a number of elements connected by a number of nodes. Each node is associated with a chosen local basis function. The unknown transmissivity distribution is then approximated by a linear combination of the basis functions, where the parameter dimension is equal to the number of unknown nodal transmissivity values (DiStefano and Rath, 1975; Yoon and Yeh, 1976; Yeh and Yoon, 1981):

$$T_e = \sum_j T_j \phi_j^e(x, y) \tag{22}$$

where the basis function ϕ_j^e is chosen in such a way that it equals 1 at the particular node j and 0 at all other nodes on the element (e). Other interpolation methods for approximating the transmissivity distribution include simple polynomial approximation, cubic spline, and kriging.

Stochastic Method

In this approach, the unknown parameter is treated as a random field, characterized by its first two moments, the mean (or drift) and the covariance function. A common approach is to assume that the logarithm of the hydraulic conductivity, $Y = \log K$, is

normally distributed (Freeze, 1975; Hoeksema, 1985a). Also, the random field is represented by a constant mean and an isotropic, exponential covariance (Dagan, 1985; Hoeksema and Kitanidis, 1985b; Wagner and Gorelick, 1989):

$$E(Y) = \mu_Y \tag{23}$$

$$\mathrm{Cov}_{YY}(x_i, x_j) = \sigma_Y^2 \exp\left(-\frac{d_{ij}}{l_Y}\right) \tag{24}$$

where σ_Y^2 = log hydraulic conductivity variance
l_Y = log hydraulic conductivity correlation scale
d_{ij} = distance between points x_i and x_j

The hydraulic conductivity can thus be estimated by identifying the three statistical parameters μ_Y, σ_Y^2, and l_Y. In this approach, overparameterization is generally avoided, and the inverse solution obtained by the maximum-likelihood estimate and cokriging is highly stable.

In addition to the traditional approaches for parameterization mentioned above, Sun et al. (1995) suggested a geological parameterization method in which the unknown parameter (hydraulic conductivity) is directly related to the geological materials, and the geological structure of the aquifer is determined by the geostatistical method of kriging.

10 PARAMETER UNCERTAINTY, PARAMETER STRUCTURE, AND OPTIMUM PARAMETER DIMENSION

Parameter identification in a distributed-parameter system should, in principle, include the determination of both the parameter structure and its value. If zonation is used to parameterize the unknown parameters, the zonation pattern (parameter structure) is represented by the number and shape of zones. On the other hand, if the finite-element method is used for parameterization, parameter structure is represented by the number and location of nodal values of parameters.

Identifying parameter structure is much more difficult than identifying parameter values for a given parameter structure. In the past three decades, only a few studies have contributed to this topic. The question of how to determine an appropriate zonation pattern was first considered by Emsellem and de Marsily (1971), who suggested that the number of zones be gradually increased until model fit no longer improved. This approach ignores the reliability of the estimated parameters. Yeh and Yoon (1976) were the first to consider both the error in model fitting and the error associated with parameter uncertainty in determining zonation pattern; to determine if a particular zone should be subdivided into smaller zones, they used the variance of the estimation error. Sun and Yeh (1985) proposed a systematic approach that can automatically identify the optimal pattern of parameter structure and its corresponding parameter values by solving a combinatorial optimization

problem. They clearly pointed out that the identified parameter values vary with the parameter structure. As a consequence, if the parameter structure is incorrect, the identified parameter values will also be incorrect. In Carrera and Neuman (1986), the dimension of parameterization is determined by Akaike's information criteria (Akaike, 1972); these criteria can also be used to compare different zonation patterns. Bellout (1992) considered the stability of pattern identification from a mathematical analysis. Recently, Zheng and Wang (1996) used the tabu search (TS) method to find the optimal zonation structure for one-dimensional problems. Eppstein and Doupherty (1996) presented an extended Kalman filter for simultaneously estimating transmissivity values and zonation pattern. A general formulation of the inverse problem that incorporates the identification of parameter structure and its parameter values is given in Sun et al. (1998). To estimate the parameter structure, some authors have attempted to incorporate directly into the solution of the inverse problems the geological structure information obtained from well-logs and seismic measurements (Rubin et al., 1992, Sun et al., 1995; Hyndman and Gorelick, 1996; Koltermann and Gorelick, 1996).

Shah et al. (1978) showed the relationship between the optimal dimension of parameterization and observations in considerable depth. The necessity to limit the dimension of parameterization has been further studied by Yeh and Yoon (1981), Yeh, et al. (1983), and Kitanidis and Vomvoris (1983). The dimension of parameterization is directly related to the quantity and quality of data (observations). In practice, the number of observations is limited and observations are corrupted with noise. Without controlling parameter dimension, instability in the inverse solution often results (Yakowitz and Duckstein, 1980). If instability occurs, parameter values will become unreasonably small (sometimes negative, which is physically impossible) and/or large, if parameter values are not properly constrained. In the constrained minimization, instability is characterized by the fact that during the inverse solution process parameter values are bouncing back and forth between the upper and lower bounds. Reduction of parameter dimension can make the inverse solution stable. As the number of zones (in the zonation case) is increased, the modeling error (least squares) decreases while the parameter uncertainty error at some point will start to increase (Shah et al., 1978; Yeh and Yoon, 1981). A trade-off of these two types of errors can then be made, from which an optimum parameter dimension can be determined. A standard procedure is to gradually increase the parameter dimension, starting from the lowest dimension, i.e., the homogeneous case, and calculate the two types of errors for each parameterization. The error in parameter uncertainty can be represented by a norm of the covariance matrix of the estimated parameters (Yeh and Yoon, 1976; Shah et al., 1978).

An approximation of the covariance matrix of the estimated parameters in nonlinear regression can be represented by the following form (Bard, 1974; Yeh and Yoon, 1976, 1981; Shah et al., 1978; Yeh, 1986):

$$\mathrm{Cov}(\hat{\mathbf{T}}) = \frac{J(\hat{\mathbf{T}})}{M - L}[\mathbf{A}(\mathbf{T})]^{-1} \tag{25}$$

where $\mathbf{J}(\hat{\mathbf{T}})$ = least-squares error
M = number of observations
L = parameter dimension
$\mathbf{A} = [\mathbf{J}_D^T \mathbf{J}_D]$
$\mathbf{J}_D$ = Jacobian matrix of $\mathbf{h}$ with respect to $\mathbf{T}$

A norm of the covariance matrix has been used to represent the error in parameter uncertainty. Norms, such as trace, spectral radius (maximum eigenvalue), and determinant have been used in the literature. Equation (25) also assumes homoscedasticity and uncorrelated errors. This assumption is generally not satisfied and the actual covariance may be much higher than given by Eq. (25).

The covariance matrix of the estimated parameters also provides information regarding the reliability of each of the estimated parameters. A well-estimtated parameter is generally characterized by a small variance as compared to an insensitive parameter that is associated with a large variance. By definition, the correlation matrix of the estimated parameter is

$$\mathbf{R} = \begin{bmatrix} \dfrac{c_{11}}{(c_{11}c_{11})^{1/2}} & \cdots & \dfrac{c_{1L}}{(c_{11}c_{LL})^{1/2}} \\ \vdots & & \vdots \\ \dfrac{c_{L1}}{(c_{LL}c_{11})^{1/2}} & \cdots & \dfrac{c_{LL}}{(c_{LL}c_{LL})^{1/2}} \end{bmatrix} \tag{26}$$

where c_{ij}'s are elements of the covariance matrix of the estimated parameters. The more sensitive the parameter, the closer and quicker the parameter will converge. A correlation analysis of the estimated parameters would indicate the degree of interdependence among the parameters with respect to the objective function. Correlation of parameters is called the *collinearity* problem. Such problems can cause slow rate of convergence in minimization and in most cases result in nonoptimal parameter estimates. A more rigorous treatment of the collinearity problem is to use the more sophisticated statistical techniques, such as ridge regression (Cooley, 1977) and the method of principal components.

11 MODEL STRUCTURE ERROR (PARAMETER STRUCTURE ERROR)

Sun et al. (1998) presented a procedure whereby the model structure error of using one model structure to replace another model structure is defined by a max–min problem that is based on the distance between the two models measured in the parameter, observation, and prediction/management space. Parameter structure error resulting from a different level of parameterization is a special case of model structure error. Without losing generality, we will use parameter structure error to represent model structure error.

The parameter structure error, $\mathrm{SE}(G_A, G_B)$, of using parameter structure G_B to replace parameter structure G_A can be defined by the following max–min problem (Sun, 1994a, 1996):

$$\mathrm{SE}(G_A, G_B) = \max_{p_A} \min_{p_B} d(G_A, \mathbf{p}_A; G_B, \mathbf{p}_B) \tag{27}$$

where d is the distance (to be defined later) between the two models, $M_A(G_A, \mathbf{p}_A)$ and $M_B(G_B, \mathbf{p}_B)$; and parameters $\mathbf{p}_A$ and $\mathbf{p}_B$ must be in their admissible regions P_A and P_B. In general, $\mathrm{SE}(G_A, G_B) \neq \mathrm{SE}(G_B, G_A)$. When G_A is a simplification of G_B, we have $\mathrm{SE}(G_A, G_B) = 0$.

The distance between the two models, $M_A(G_A, \mathbf{p}_A)$ and $M_B(G_B, \mathbf{p}_B)$, as generalized by Sun et al. (1998), can be defined as:

$$d(M_A, M_B) = d_E + \mu d_D + \lambda d_P \tag{28}$$

where

$$d_E(M_A, M_B) = \|g_E(M_A) - g_E(M_B)\|_E \tag{29}$$

$$d_D(M_A, M_B) = \|h_D(M_A) - h_D(M_B)\|_D \tag{30}$$

$$d_P(M_A, M_B) = \|\bar{\mathbf{p}}A - \bar{\mathbf{p}}_B\|_{\bar{G}} \tag{31}$$

where subscript E denotes a prediction/management alternative and its associated prediction space; $\|\cdot\|_E$ is a norm defined in the prediction space; subscript D denotes an observation design and its associated observation space; $h_D(M_A)$ and $h_D(M_B)$ are "observations" based on the same observation design but generated from difference models, M_A and M_B; $\|\cdot\|_D$ is a norm defined in the observation space; $\bar{G}$ is a parameter space having a common overparameterization structure of G_A and G_B; $\bar{\mathbf{p}}_A$ and $\bar{\mathbf{p}}_B$ are spans of $\mathbf{p}_A$ and $\mathbf{p}_B$; $\|\cdot\|_{\bar{G}}$ is a norm defined in $\bar{G}$; μ and λ are weighting coefficients. It is clear that by varying the weighting coefficients, one can emphasize the importance of each distance in the parameter, observation, or prediction/man-management space. As a result, this will influence the inverse solution.

12 GENERALIZED INVERSE PROCEDURE

The generalized inverse procedure seeks to minimize the weighted composite objective function as represented by Eq. (28). In this procedure, the unknown model structure (parameter structure) and its corresponding parameter values are determined not only from prior information and observations but also by the accuracy requirement in model applications. Sun et al. (1997, 1998) presented a stepwise regression procedure for a simultaneous estimation of parameter structure and its corresponding parameter values. The procedure starts from a homogeneous parameter structure and gradually increases the complexity of the parameter structure. For a given set of data and a specified model reliability requirement, the method, at

each level of complexity, calculates both the least-squares error as well as the parameterization error of using a simpler parameter structure to replace a more complex parameter structure. The method is most general as it considers errors in the parameter, observation, as well as prediction/management space. The established procedure allows one to determine whether a more complex parameter structure is needed or to conclude that data are insufficient to meet the specified model reliability requirement; and hence, additional data are needed.

In this procedure, we form a series of parameter structures of increasing complexity:

$$G_1, G_2, G_3, \ldots, G_m, \ldots$$

where G_1 represents a homogeneous structure, G_2 a two-zone structure, and so forth; G_2 is generated from G_1 by dividing it into two zones and, generally, G_{m+1} is generated from G_m by dividing one of the zones of G_m into two zones. At each level, we calculate the residual error (RE) and the parameter structure error (SE). Specifically, the following steps are involved:

Step 1. Let G_1 be a homogeneous parameter structure, we solve the generalized inverse problem to find $\mathbf{p}_1^*$ and the corresponding residual error RE_1. In general, RE can be found by minimizing a linear combination of the norms in the parameter and observation space. Details can be found in Sun et al. (1998).

Step 2. Divide G_1 into two zones to generate G_2. The method suggested by Sun and Yeh (1995) can be used to optimize simultaneously the zonation pattern and its corresponding parameter values. In this step, we find a model $M_2(G_2, \mathbf{p}_2^*)$ and its residual error RE_2. RE_2 must be smaller than RE_1 because a homogeneous parameter structure is being replaced by a two-zone structure to fit the same set of observations.

Step 3. Calculate the parameter structure error SE_1 by using G_1 to replace G_2. Details with regard to how to calculate SE_1 are presented in Sun et al. (1998).

Step 4. If both SE_1 and RE_2 are large, we continue to increase the parameter structure complexity by finding the optimum three-zone parameter structure M_3 $(G_3, \mathbf{p}_3^*)$ and RE_3.

Step 5. Calculate the parameter structure error SE_2 of using G_2 to replace G_3. If both SE_2 and RE_2 are large, we repeat steps 4 and 5 to obtain $M_4(G_4, \mathbf{p}_4^*)$, RE_4 and SE_3, and so forth. Assume that through this procedure we have found $M_{m+1}(G_{m+1}, \mathbf{p}_{m+1}^*)$, RE_{m+1}, and SE_m.

Step 6. Then consider the following four cases:

1. If both RE_{m+1} and SE_m are large compared to the observation error and accuracy requirement of the prediction/management problem, respectively, increase m by one and repeat step 5.
2. If both RE_{m+1} and SE_m are small, stop and use M_{m+1} as the identified model.
3. If RE_{m+1} is large but SE_m is small, either stop or continue to increase the complexity until RE_{m+1} becomes small;
4. If RE_{m+1} is small but SE_m is large, additional data are required.

In case 1, the identified model cannot satisfy the accuracy requirement of the given model application but, at the same time, the existing data still have the potential to provide more information. Therefore, we increase the parameter structure complexity. In cases 2 and 3, when the complexity of the parameter structure is increased, the prediction/management solution is not significantly improved; thus, we can either accept the identified model or if existing data still contain more information, we can continue to increase the parameter structure complexity. In case 4, the information contained in the existing data is insufficient to identify a reliable model and, thus, additional data are required to be collected.

13 CONCLUSIONS

The development of groundwater simulation models in the early 1970s provided groundwater planners with quantitative techniques for analyzing alternative groundwater pumping or recharge strategies. The accuracy of a simulation model is dependent, to a certain extent, on the accuracy of the inverse solution, which in turn is determined by the quantity and quality of data. The inverse problem is inherently nonunique and unstable. It has been well understood that the number of unknown parameters must be reduced to obtain a unique and stable solution of the inverse problem. The reduction of the number of unknown parameters is achieved by means of parameterization. It also has become apparent that parameterization and its corresponding parameter values are interdependent and must be estimated simultaneously.

A recent advancement made in groundwater modeling has been the development of a generalized inverse procedure. This procedure allows us to analyze the errors in the parameter, observation, and prediction/management space. The requirement of finding the true parameter values in the classical inverse problem is replaced by a weaker requirement. More importantly, it helps us resolve the following two problems: (1) How complex should a groundwater model structure be for a given model application? (2) Are existing data sufficient for developing a reliable model for the stipulated prediction/management objective? The generalized inverse procedure attempts to find the simplest model structure for a given model application. Such a model requires the minimum amount of data to calibrate.

REFERENCES

Akaike, H., Information theory and an extension of the maximum likelihood principle, in B.N. Petrov and F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information Theory, Supplement to Problems of Control and Information Theory*, Akad. Kiado, Budapest, 1972, pp. 267–281.

Anderson, M. P., and W. W. Woessner, *Applied Groundwater Modeling: Simulation of Flow and Advective Transport*, Academic, San Diego, CA, 1992.

Bard, Y., *Nonlinear Parameter Estimation*, Wiley, New York, 1974.

Bear, J., *Hydraulics of Groundwater*, McGraw-Hill, New York, 1979.

Bedient, P. B., H. S. Rifai, and C. J. Newell, *Ground Water Contamination*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

Bellout, H., Stability result for the inverse transmissivity problem, *J. Math. Anal. Applicat.*, *168*, 13–27, 1992.

Bredehoeft, J. D., and G. F. Pinder, Digital analysis of areal flow in mutiaquifer groundwater systems: A quasi-three dimensional model, *Water Resour. Res.*, *6*(3), 885–888, 1980.

Brutsaert, W., and H. A. Ibrahim, On the first and second linearization of the Boussinesq equation, *J. Am. Soc. Geophys.*, *11*, 549–554, 1966.

Carrera, J., State of the art of the inverse problem applied to flow and solute transport equations, in E. Custodio, A. Gurgui, and J. P. Lobo Ferreira (Eds.), *Groundwater Flow and Quality Modeling*, D. Reidel, Hingham, MA, 1988, pp. 549–583.

Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under transient and steady state conditions, 3, Application to synthetic and field data, *Water Resour. Res.*, *22*(2), 228–242, 1986.

Chapman, M. J., and K. R. Godfrey, On structual equivalence and identifiability constraint ordering, in E. Walter (Ed.), *Identifiability of Parametric Models*, Pergamon, New York, 1987, pp. 32–39.

Chavent, G., M. Dupuy, and P. Lemonnier, History matching by use of optimal theory, *Soc. Pet. Eng. J.*, *15*(1), 74–86, 1975.

Coats, K. H., J. R. Dempsey, and J. H. Henderson, A new technique for determining reservoir description from filed performance data, *Soc. Pet. Eng. J.*, *10*(1), 66–74, 1970.

Coleman, T. F., A note on New Algorithms for constrained minmax optimization, *Math. Programming*, *15*, 239–242, 1978.

Cooley, R. L., A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 1. Theory and numerical properties, *Water Resour. Res.*, *13*(2), 318–324, 1977.

Cooley, R. L., A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 2. Application of statistical analysis, *Water Resour. Res.*, *15*(3), 603–617, 1979.

Cooley, R. L., Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 1. Theory, *Water Resour. Res.*, *18*(4), 965–976, 1982.

Dagan, G., Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem, *Water Resour. Res.*, *21*(1), 65–72, 1985.

DeWiest, R. J. M., *Geohydrology*, Wiley, New York, 1965.

DiStefano, N., and A. Rath, An identification approach to subsurface hydrological systems, *Water Resour. Res.*, *11*(6), 1005–1012, 1975.

Eagleson, P. S., *Dynamic Hydrology*, McGraw-Hill, New York, 1970.

Emsellem, Y., and G. de Marsily, An automatic solution for the inverse problem, *Water Resour. Res.*, *7*(5), 1264–1283, 1971.

Eppstein, M. J., and D. E. Dougherty, Simultaneous estimation of transmissivity values and zonation, *Water Resour. Res.*, *32*(11), 3321–3336, 1996.

Ezzedine, S., and Y. Rubin, A geostatistical approach to conditional estimation of spatially distributed solute concentration and notes on the use of tracer data in the inverse problem, *Water Resour. Res.*, *32*(4), 853–862, 1987.

Freeze, R. A., A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media, *Water Resour. Res.*, *11*(5), 725–741, 1975.

Freeze, R. A., and J. A. Cherry, *Groundwater*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

Haber, R., and H. Unbehauen, Structure identification of nonlinear dynamic system—A survey on Input/Output approaches, *Automatica*, *26*(4), 651–677, 1990.

Hantush, M. S., Nonsteady flow to flowing wells in leaky aquifer, *J. Geophys. Res.*, *64*, 1943–1052, 1959.

Hill, M. C., *A Computer Program (MODFLOWP) for Estimating Parameters of a Transient, Three-Dimensional, Ground-Water Flow Model Using Nonlinear Regression*, Open-File Report 91-484, U.S. Geological Survey, Denver, CO, 1992.

Hoeksema, R. J., and P. K. Kitanidis, Analysis of the spatial structure of properties of selected aquifers, *Water Resour. Res.*, *21*(4), 563–572, 1985a.

Hoeksema, R. J., and P. K. Kitanidis, Comparison of Gaussian conditional mean and Kriging estimation in the geostatistical solution of inverse problem, *Water Resour. Res.*, *21*(6), 825–836, 1985b.

Hyndman, D. W., and S. M. Gorelick, Estimating lithologic and transport properties in three dimensions using seismic and tracer data: The Kesterson aquifer, *Water Resour. Res.*, *32*(9), 2659–2670, 1996.

Kitanidis, P., Quasi-linear geostatistical theory for inversing, *Water Resour. Res.*, *31*(10), 2411–2420, 1995.

Kitanidis, P. K., and E. G. Vomvoris, A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, *Water Resour. Res.*, *19*(3), 677–690, 1983.

Koltermann, C. E., and S. M. Gorelick, Heterogeneity in sedimentary: A review of structure-imitating, process-imitating, and descriptive approaches, *Water Resour. Res.*, *32*(9), 2617–2658, 1996.

Loaiciga, H. A., and M. A. Marino, The inverse problem for confined flow: Identification and estimation with extension, *Water Resour. Res.*, *23*(1), 92–104, 1987.

Loaiciga, H. A., R. B. Laipnik, P. K. Herdak, and M. A. Marino, Effective hydraulic conductivity of nonstationary aquifers, *Stochast. Hydrol. Hydraul.*, *8*(1), 1–17, 1994.

McDonald, M. G., and A. W. Harbaugh, *A Modular Three-Dimensional Finite Difference Ground-Water Flow Model*, Open-File Report 83-875, U.S. Geological Survey, Denver, CO, 1988.

McLaughlin, D., and L. R. Townley, A reassessment of the groundwater inverse problem, *Water Resour. Res.*, *32*(5), 1131–1162, 1996.

Neuman, S. P., Calibration of distributed parameter groundwater flow models viewed as a multiple-objective decision process under uncertainty, *Water Resour. Res.*, *9*(4), 1006–1021, 1973.

Prickett, T. A., and C. O. Lonnquist, *Selected Digital Computer Techniques for Groundwater Resource Evaluation*, Bulletin No. 55, Illinois State Water Survey, Urbana, IL, 1971.

Poeter, E. P., and M. C. Hill, Inverse models: a necessary next step in ground-water modeling, *Ground Water*, *35*(2), 250–260, 1997.

RamaRoa, B. S., M. A. LaVenue, G. de Marsily, and M. G. Marietta, Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields, 1, Theory and computational experiments, *Water Resour. Res.*, *31*(3), 475–493, 1995.

Rubin, Y., and G. Dagan, Stochastic identification of transmissivity and effective recharge in steady groundwater flow, 1, Theory, *Water Resour. Res.*, *23*(7), 1809–1916, 1992.

Rubin, Y., G. Mavko, and J. Harris, Mapping permeability in heterogeneous aquifers using hydrological and seismic data, *Water Resour. Res.*, *28*(7), 1809–1816, 1992.

Rustem, B., A constrained min-max algorithm for rival models of the same economic system, *Math. Programming*, *53*, 279–295, 1992.

Shah, P. C., G. R. Gavalas, and J. H. Seinfeld, Error analysis in history matching: The optimum level of parameterization, *Soc. Pet. Eng. J.*, *18*(3), 219–228, 1978.

Sun, N-Z., *Inverse Problems in Groundwater Modeling*, Kluwer Academic, Norwell, Mass., 1994a.

Sun, N-Z., *Mathematical Modeling of Groundwater Pollution*, Springer-Verlag, New York, 1994b.

Sun, N-Z., Identification and reduction of model structure for modeling distributed parameter systems, in J. Gottlieb and P. DuChateau (Eds.), *Parameter Identification and Inverse Problems in Hydrology, Geology and Ecology*, Kluwer Academic, Norwell, Mass., 1996.

Sun, N-Z., and W. W-G. Yeh, Identification of parameter structure in groundwater inverse problem, *Water Resour. Res.*, *21*(6), 869–883, 1985.

Sun, N-Z., and W. W-G. Yeh, Coupled inverse problem in groundwater modeling, 1, Sensitivity analysis and parameter identification, *Water Resour. Res.*, *26*(10), 2507–2525, 1990.

Sun, N-Z., and W. W-G. Yeh, A stochastic inverse solution for transient groundwater flow: Parameter identification and reliability analysis, *Water Resour. Res.*, *28*(12), 3269–3280, 1992.

Sun, N-Z., M-C. Jeng, and W. W-G. Yeh, A proposed geological parameterization method for parameter identification in three-dimensional groundwater modeling, *Water Resour. Res.*, *31*(1), 89–102, 1995.

Sun, N-Z., M-C. Jeng, and W. W-G. Yeh, Model structure identification: The generalized inverse problem, in K. W. Watson and Z. Zaporozec (Eds.), *Advances in Ground-Water Hydrology*, American Institute of Hydrology, Tampa, FL, 1997, pp. 130–134).

Sun, N-Z., S. Yang, and W. W-G. Yeh, A proposed stepwise regression method for model structure identification, *Water Resour. Res.*, *34*(10), 2561–2572, 1998.

Todd, D. K., *Groundwater Hydrology*, Wiley, New York, 1980.

Wagner, B. J., and S. M. Gerelick, Reliable aquifer remediation in the presence of spatially variable hydraulic conductivity: From data to design, *Water Resour. Res.*, *25*(10), 2211–2225, 1989.

Willis, R., and W. W-G. Yeh, *Groundwater Systems Planning and Management*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

Xiang, Y., J. F. Sykes, and N. R. Thomson, A composite L_1 parameter estimator for model fitting in groundwater flow and solute transport simulation, *Water Resour. Res.*, *29*(6), 1661–1674, 1993.

Yakowitz, S., and L. Duckstein, Instability in aquifer identification: Theory and case studies, *Water Resour. Res.*, *16*(6), 1045-1064, 1980.

Yeh, W. W-G., Review of parameter identification procedure in groundwater hydrology: The inverse problem, *Water Resour. Res.*, *22*(2), 9–108, 1986.

Yeh, W. W-G., Systems analysis in ground-water planning and management, *J. Water Resour. Planning and Mgmt.*, *118*(3), 224–237, 1992.

Yeh, W. W-G., and N-Z. Sun, An extended identifiability in aquifer parameter identification and optimal pumping test design, *Water Resour. Res.*, *20*(12), 1837–1847, 1984.

Yeh, W. W-G., and N-Z. Sun, Variational sensibility analysis, data requirements, and parameter identification in a leaky aquifer system, *Water Resour. Res.*, *26*(9), 1827–1938, 1990.

Yeh, W. W-G., and Y. S. Yoon, A systematic optimization procedure for the identification of inhomogeneous aquifer parameters, in Z. A. Saleen (Ed.), *Advances in Groundwater Hydrology*, American Water Resources Association, Minneapolis, Minnesota, 1976, pp. 72–82.

Yeh, W. W-G., and Y. S. Yoon, Aquifer parameter identification with optimum dimension in parameterization, *Water Resour. Res.*, *17*(3), 664–672, 1981.

Yeh, J. T-C., and J. Zhang, A geostatistical inverse method for variably saturated flow in the vadose zone, *Water Resour. Res.*, *32*(9), 2757–2766, 1996.

Yeh, W. W-G., Y. S. Yoon, and K. S. Lee, Aquifer parameter identification with Kriging and optimum parameterization, *Water Resour. Res.*, *19*(1), 225–233, 1983.

Yoon, Y. S., Yeh, W. W-G. Parameter identification in an inhomogeneous medium with the finite-element method, *Soc. Petrol. Eng. J.*, *16*(4), 217–226, 1976.

Zheng, C., and P. Wang, Parameter structure identification using tabu search and simulated annealing, *Adv. Water Resour.*, *19*(4), 215–224, 1996.

CHAPTER 29

SURFACE RUNOFF GENERATION

KEITH BEVEN

1 INTRODUCTION: DEFINING RUNOFF

There are a number of different definitions of runoff that have been used either explicitly or implicitly in hydrological analyses over the years. In what follows we will use a working definition that *runoff* is that part of the rainfall falling on a catchment area that eventually leaves the catchment as a surface streamflow, whatever the flow pathway that the water has followed on its way to the stream channel. Thus this definition includes both surface and subsurface runoff pathways. Dunne (1978) provides a review of field studies of surface and subsurface runoff generation processes that remains one of the best summaries available.

For a long time, following the work of Robert Horton in the 1930s, storm runoff was often taken to be equivalent to a purely surface runoff process. Horton suggested that the soil surface acted as a separating surface, between fast (surface) storm runoff processes and slow (subsurface) flow processes contributing to *baseflow* (Horton, 1933). This concept still underlies a great deal of hydrological analysis even though we now know that this is often not the case and that much of the runoff in a stream channel seen during a storm event may have followed subsurface flow pathways. C. R. Hursh, working at the same time as Horton, was demonstrating the importance of subsurface flow in storm hydrograph generation in the Coweeta catchments in North Carolina (e.g., Hursh, 1944).

The best basis for the analysis of runoff in a catchment is to allow that there may be a spectrum of surface and subsurface flow velocities and path lengths, which must be expected to change with the state of wetness of the catchment area and with the nature and spatial pattern of a rainfall event. In some conditions or locations, subsurface flow processes may dominate runoff generation; in other conditions or locations (even within the same catchment), surface flow process may dominate. Indeed, in

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

some conditions and locations, subsurface flow may saturate the soil and return to the surface as a *return flow*, the area of saturation also acting as a *dynamic contributing area* for surface runoff due to additional rainfall inputs, that will expand and contract as the catchment wets and dries (Fig. 1). Thus, it is only for convenience that, in what follows, we treat subsurface and surface runoff processes in turn.

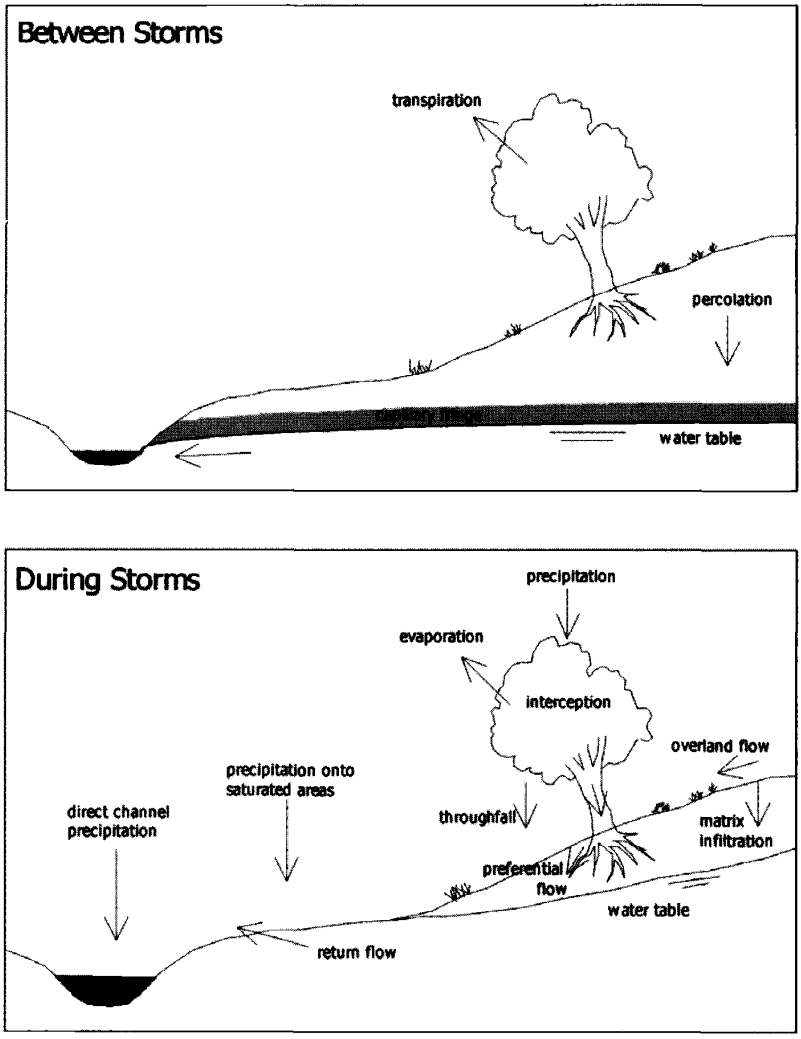


Figure 1 Processses of hillslope response to rainfall (after Beven, 2001).

2 GENERATION OF SUBSURFACE RUNOFF

Role of Soil in Runoff Processes

When rainfall falls on the land surface, in many environments the first thing it hits will be a vegetation canopy. This has the effect of retaining some of the rain on leaf surfaces as *interception* and redistributing the rest down to the ground surface as *throughfall* and *stemflow*. At the ground surface, therefore, the rate of supply of water will no longer be spatially uniform. There may be local concentrations at the base of stems or trunks; there may be other areas that receive lower intensity inputs.

This distribution of intensities falls onto a surface that will have a variable infiltration capacity due to variations in soil properties and initial moisture status. In particular, the properties at the soil surface will be very important in controlling how much of the rainfall can infiltrate into the soil. However, in many environments much of the rainfall input will infiltrate into the soil unless the soil is already completely saturated. It will do so by one of two routes, by direct infiltration into the soil matrix and by infiltration into pathways due to larger structural voids (cracks, root channels, animal or insect burrows, etc). The latter may be very important since the faster flow velocities associated with structural porosity can mean that the water can move rapidly into the soil, bypassing parts of the soil matrix as a *preferential flow* (see, e.g., Beven and Germann, 1982). Some of this water will be absorbed into the soil matrix at depth, some may move rapidly downslope within the preferential flow pathways, and some may be moved rapidly vertically to the local water table. Rates of movement will depend on the input rainfall intensity, the permeability and initial moisture content of the soil matrix, the structural characteristics of the soil, and the depth to the water table. It is well known from soil thin sections, however, that water movement in structural pathways can lead to transport of clay particles that are deposited in thin layers called *cutans*, which can then restrict the infiltration of water into the soil matrix and prolong the preferential flow. It is also known that preferential flow can be important in the transport of contaminants, such as pesticides and herbicides, in runoff since such contaminants are often sorbed to fine particles.

Within the soil matrix, water movement into soil that is not fully saturated will take place primarily vertically as a *wetting front*. The propagation of the wetting front will again depend on the antecedent wetness of the soil, the input intensity, and the matrix hydraulic characteristics. Flow within a continuous soil matrix can usually be described by the *Richards equation*, which is based on the unsaturated form of *Darcy's law*. The Richards equation is difficult to solve for general situations of practical hydrological interest, but there are now very many approximate numerical solution codes available (see Chapter 28). All such solution algorithms will require the specification of the soil hydraulic characteristics, which will depend on the texture and organic matter content of the matrix.

Based on many thousands of measurements, empirical relationships have been developed between the easily measured texture characteristics and the much more difficult to measure soil hydraulic properties (see, e.g., Rawls and Brakensiek,

1989). These empirical regressions are often useful but must be used with care. The estimates obtained in this way are subject to significant estimation error and are also only as good as the original data on which they are based. In this case the measurements were usually based on small soil samples and did not include any effects of the structural porosity. In any case, preferential flows may not be well described by the Richards equation, and it has proven very difficult to develop a comprehensive description of preferential flow with parameters that can easily be estimated in applications.

A useful simple analogy for the movement of water in both matrix and preferential flow pathways is that of the wetting front as a kinematic shock or locally pistonlike front, in which the rate of propagation of the front is given as

$$c = I/\Delta\theta$$

where c is the velocity of the front, I is the local input intensity, and $\Delta\theta$ is an effective change in moisture content across the front. This is an approximation, applicable only where the input rate I does not exceed the infiltration capacity of the soil at any depth, but it is then readily seen that the wave speed c increases with the input intensity and decreases with the change in moisture content. Thus, if $\Delta\theta$ is small, either in a preferential flow pathway (ignoring losses due to sorption into the matrix) or because the soil matrix is already wet, the front may move quickly into the soil. For a low input intensity I, infiltrating into a dry soil with an effectively high $\Delta\theta$, the speed of the wetting front may be very much lower. With a variation of effective intensities at the ground surface, and a variety of effective local $\Delta\theta$ values in different flow pathways, there will be a distribution of wetting front velocities in the soil.

Some of the infiltrating water will be retained in the soil and later evaporated or transpired back to the atmosphere (see Chapter 26), but some of the wetting resulting from a storm rainfall may reach an existing water table, or will induce saturation at the base of the soil profile, or a *perched* zone of saturation above a horizon of lower permeability in the soil profile. The wetting or *recharge* will induce a response in the saturated zone that will ultimately produce some subsurface runoff.

Recharge and Downslope Flow in a Saturated Zone

In fully saturated soil the propagation of the effects of changes in the boundary conditions, such as those due to recharge, is much more rapid than in the unsaturated zone. In shallow subsurface systems, such as where a shallow soil overlays an impermeable rock bed, most of the downslope flow toward stream channels will take place in the saturated zone. Because of the more rapid dissipation of local pressure differences in the saturated zone, a description of flow processes based on Darcy's law is generally more acceptable, even if preferential flow pathways are still contributing to the flow, since those pathways will be subject to similar pressure gradient conditions to the saturated matrix (with the reservation that in large

pipe systems the flow may be turbulent and transitional rather than laminar and Darcy's law will not be valid).

Again, in relatively shallow soil systems a kinematic description is a useful analogy for the saturated zone, if we can assume that the local hydraulic gradient is approximately equal to the local slope angle (or even better the local bedslope angle). In this case, the equation of flow is the kinematic wave equation

$$W_x \phi_e \frac{\partial h}{\partial t} = T_h \sin \alpha \frac{\partial W_x h}{\partial x} + r_{x,t}$$

where h is the depth of saturation above the bed, x is distance measured along the slope, W_x, is the width of the slope at point x, α is the local slope angle, r is the recharge rate at point x and time t, and T_h is the integral of the saturated soil hydraulic conductivity function K_h over the depth of saturation h and is called the *transmissivity*; T_h and K_h will be a function of h that may be nonlinear for many soil profiles. The local downslope Darcian velocity of flow (volume flux per unit cross-sectional area) is then given by

$$V = K_h \sin \alpha$$

The mean pore water velocity is then given by

$$V_p = K_h \sin \alpha / \phi$$

where ϕ is the porosity of the soil. This is the mean velocity of the water itself. The kinematic wave velocity is given by

$$V_c = Kh \sin \alpha / \phi_c$$

where ϕ_c is an effective storage coefficient for the soil, or effectively in this case, the difference between the soil water content just above the water table and saturation. This last velocity is the rate at which disturbances to the flow are propagated in the direction of flow. The effective storage coefficient will generally be much less than the porosity of the soil, particularly in soils that are near saturation above the water table. Thus the wave speed may be very much faster than the mean pore water velocity, which will be faster than the Darcian velocity (since the porosity must itself be less than 1). The implications of this are that the effects of a recharge to the water table will move downslope much faster than the speed at which the water itself is moving. This will, in general, cause a rise in the subsurface outflow into the downslope stream channel more quickly than the inputs can flow toward that channel. This is one reason why in humid environments subsurface flow processes can make more rapid contributions to storm runoff than has been generally accepted in the past.

Again, Eq. (2), the kinematic wave equation, is an approximation. A fuller description will allow for fully three-dimensional flows in both soil and bedrock,

perhaps including flow through bedrock fractures, with time-variable hydraulic gradients (see, e.g., Rasmussen et al., 2000). The same behavior of pressure wave transmission being faster than pore water velocities being in turn faster than Darcian velocities will hold. This is important in understanding the results of tracer experiments.

Equation (2) has been written in a way that allows for the width of the hillslope to vary downslope. It has been known for some time that an important control on the production of subsurface runoff, and of the occurrence of saturated dynamic contributing areas, is the form of hillslopes, in terms of both convergence or divergence in plan, and convexity or concavity in section. Soil water contents will tend to be higher and the saturated zone nearer to the surface in areas that are both convergent and concave. Such areas tend to be found particularly in *zero-order* headwater basins close to the heads of channels or the appearance of springs. These are areas where the soil is most likely to be close to saturation and consequently will show the greatest potential for acting as runoff source areas or *dynamic contributing areas* by either surface or subsurface flow processes.

Hillslope topography is, however, not the only cause of variability in flow rates. There is an increasing appreciation of the role of the geological structure of a catchment in controlling the subsurface flow pathways, even in catchments where there is no deep aquifer. The tracing experiments of Genereux et al. (1993), for example, revealed strong variability in the channel inputs in the Walker Branch catchment, Tennessee, that appeared to result from the bedding structure of the underlying rocks. Fracture systems in the near surface geology can also lead to the concentration of flow in certain locations. The occurrence of local perennial or seasonal springs is an indication that such effects may be important. Deeper fracture systems and flows along fault lines may also have an effect on subsurface flow pathways but are very difficult to study. Usually it is necessary to infer the presence of such flow pathways from the geochemical characteristics of baseflows.

For areas of relatively homogeneous shallow soils, it has been demonstrated that one way of predicting the location of such source areas is by use of the pattern of the topographic index a/s (e.g., Kirkby, 1978; see also O'Loughlin, 1981), which is the ratio of the area draining from upslope through unit contour length at any point in the catchment, to the slope angle at that point. The upslope area a represents the propensity for water to collect at a point; while the slope s represents the ease with which that point will drain. Approximate steady-state theory suggests that the index can be used as an index of hydrological similarity in that, other things being equal, points with similar values of the index should respond in a hydrologically similar way (Fig. 2). The topographic index has been incorporated into the rainfall–runoff model TOPMODEL and land surface parameterization TOPLATS, which aim to predict the dynamics of the surface and subsurface contributing areas and spatial patterns of latent heat flux in a simple way (Beven et al., 1995; Famiglietti et al., 1992). For recent critiques of the success of using the topographic index to represent contributing area dynamics see Beven (1997).

Use of the topographic index assumes that there is a consistent downslope flow of water on the hillslopes, but this is not always the case, particularly in catchment areas that are subject to extended drying periods. In such catchments, the effective subsur-

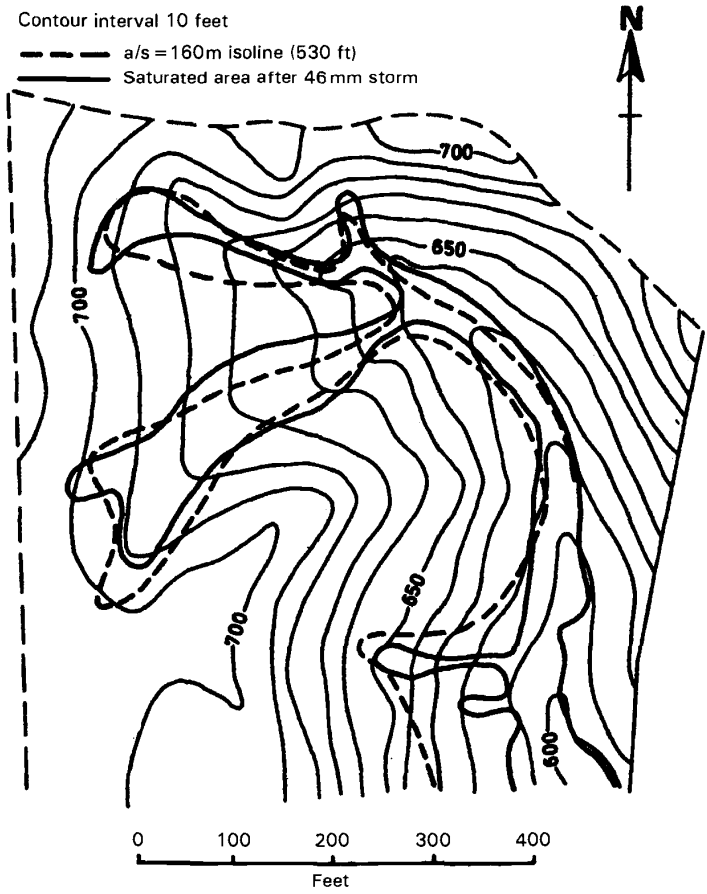


Figure 2 Pattern of the topographic index a/s in comparison with measured areas of surface saturation in basin WC-4 Sleepers River, Vermont (after Kirkby, 1978).

face contributing areas to the channel may not normally extend to the catchment divides and the development and connectivity of local saturation zones may be important. As noted earlier, in some soils, *perched* water tables may develop over a permeability break in the profile, resulting in increased downslope flow velocities without the profile being saturated to its base. This will tend to occur first in areas at the base of slopes and in hillslope hollows where the soil is normally wetter prior to an event (e.g., Weyman, 1970).

A final subsurface flow process that can lead to rapid subsurface responses is flow in natural soil pipes or along *percolines* of higher permeability (see, e.g., Beven and

Germann, 1982; McDonnell, 1990). Artificial drainage can have a similar effect, at least under wet conditions (under dry conditions artificial drainage can enhance the storage deficit of the soil prior to a storm and thereby lead to a lower runoff coefficient for an event.

Old and New Water Contributions to Storm Runoff

There have been many studies in the last 30 years that have made use of the natural geochemical characteristics of runoff to give at least an approximate separation of the storm hydrograph into a contribution from the rainstorm itself (the "new" water) and water stored in the catchment prior to the event ("old" water) (see, e.g., Sklash, 1990). Most such studies have used a two-component separation (into old and new water), which requires the assumptions that the geochemical characteristics of the two components are distinctly different and are constant in both space and time. The old water concentration is usually taken as that of the stream water baseflow component, measured prior to the start of a storm. The results, under these assumptions, have often suggested that a large proportion of the storm hydrograph may be made up of old water.

The assumptions, however, have often been questioned, particularly that of the constancy of the old water component. A number of studies have therefore introduced a third component, or *end member*, with the chemical characteristics of the "soil water," as determined from direct sampling. A separation or *end member mixing analysis* (*EMMA*) into three components requires measurements on at least two different tracers and may also be subject to some uncertainty (see, Bazemore et al., 1994; Fig. 3). The results, however, still usually suggest that a large proportion of the hydrograph is made up of water stored in the catchment prior to the event. This proportion would be reduced if it were shown that the new water rapidly takes on the tracer concentration characteristics of soil water, especially for reactive natural tracers such as silica, since it is known that equilibration times for dissolution or desorption can be short relative to storm duration in some circumstances.

Although these results have led to a reevaluation of the Hortonian concept that storm runoff is predominantly provided by rain water running off over the surface of the soil, there is no mystery about the hydrograph showing a large proportion of old water. This requires a displacement of water stored in the catchment prior to an event by the incoming rain water during the event. The simplified kinematic analysis of the different subsurface velocities presented above suggests that the wave speed will be greater than the mean pore water velocity and therefore displacement of old water into the stream would be expected. In addition, at least for moderate storms in humid catchments, the volume of water stored in the profile prior to an event may be much greater than the volume of event water (a 1-m soil profile, with an average of 25% moisture content in the profile, contains the equivalent of 250 mm of water per unit area). Thus, there will often be more than enough water available to be displaced. However, the proportion of old water might be expected to decrease as storm magnitude increases, but very few measurements have been reported for large-magnitude events.

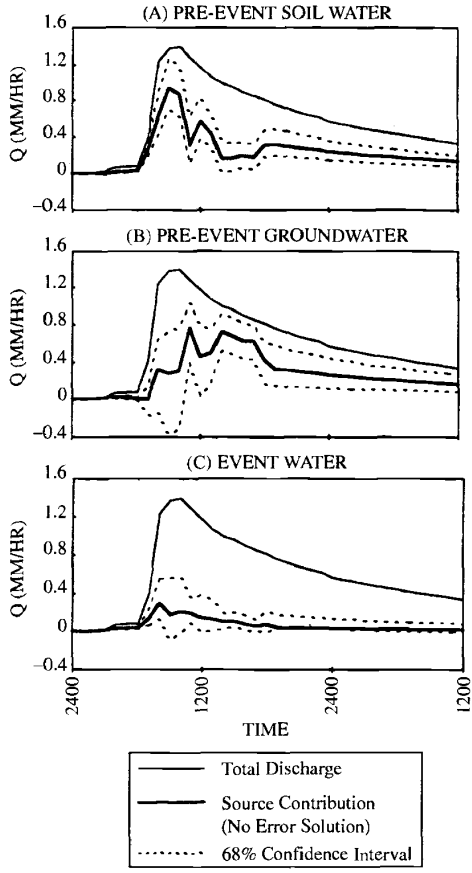


Figure 3 Separation of a stream hydrograph into rainfall, soil water, and groundwater contributions with estimates of the uncertainty associated with each component (after Bazemore et al., 1994).

Special Case of Runoff Generation by Snowmelt

Many hydrological regimes, particularly mountain regimes, are dominated by the spring snowmelt component. Snowmelt has particular characteristics in generating the snowmelt hydrograph. It tends to have only low intensities since even after the snow pack is "ripe" and ready for melt, rates of melt are limited on a daily basis by the energy available to supply the latent heat necessary to convert ice and snow to liquid water. Initially, routing through the snowpack may also diffuse the daily melt signal, and there may be some refreezing of water at night. Another interesting

feature of the melt process is that it will have a characteristic spatial pattern since, in general, south facing slopes will melt before north facing slopes (in the Northern Hemisphere) and a low elevation snowpack before a high elevation pack. There may also be spatial variations in melt associated with differences in the storage of snow as a result of drifting during the winter period.

The response of a catchment during the snowmelt period may depend very much on the state of the soil. If the soil is frozen, then it is likely that infiltration rates may be limited and there is a greater chance of the melt generating a downslope surface runoff through the base of the pack. If the soil is unfrozen, then the low intensity of the melt will usually mean that the bulk of the melt will infiltrate into the soil profile. Depending on the weather conditions prior to a pack being established, it is quite possible that in some years the soil surface remains frozen all winter, while in other years the surface is unfrozen at the start of the melt season. The responses expected during melt might then be very different in different years.

Melt rates can be greatly accelerated, if warm rain falls on a ripe snowpack, and in some parts of the world rain on snow events can be a significant cause of flooding. The rain adds both water to the event volume and heat resulting in increased rates of melt. This type of event was involved in the northern California floods of early 1996.

3 GENERATION OF SURFACE RUNOFF

Infiltration Excess Surface Runoff Generation

The classical model of surface runoff generation is by an infiltration excess mechanism in which rainfall intensity exceeds the local infiltration capacity of the soil for a sufficient period of time for any *depression storage* at the soil surface to be satisfied such that downslope flow is initiated. This will not occur where the permeability of the soil is high in comparison with expected rainfall intensities, and even when this is not the case there may be an initial period when all the rainfall infiltrates before bringing the surface to saturation (the *time to ponding*). There have been many infiltration equations proposed, either empirical or based on various approximate solutions to Darcy's law, with the aim of predicting times to ponding, infiltration capacities, and the production of surface runoff. All require the specification of some parameter values that measurements suggest may vary dramatically even within a single soil type. Spatial variability of soil characteristics may therefore be important in the production of surface runoff, since infiltration excess runoff will start first in areas of low infiltration capacity. Where runoff flows downslope onto areas of higher infiltration capacity as run-on, there may be further infiltration of part or all of the water before it reaches a stream channel.

Infiltration capacities of the soil can also be greatly enhanced by the presence of macropores [cracks, root channels, animal burrows see Beven and Germann (1982)] or reduced by the presence of surface crusting resulting from the redistribution of fine particles by raindrop impact (Römkens et al., 1990) such that the properties of the bulk soil matrix alone may not be a good indication of actual infiltration rates.

Saturation Excess Surface Runoff Generation

Surface runoff may be found in areas of high infiltration capacity soils if the rainfall falls on soil that is saturated to the surface or if return flow from the subsurface occurs onto a saturated surface. This may not, in fact, require saturation of the whole soil profile. The buildup of a perched water table, for example, due to a permeability break between horizons in the soil profile, might result in saturation to the surface and the consequent generation of surface runoff. Surface runoff produced in this way has been studied, for example, by Bonell et al. (1981).

Saturation is most likely to occur where upslope contributing areas are large, such as in valley bottoms and hillslope hollows, and where effective hydraulic gradients are small. Convergence of flow lines will tend to increase the likelihood of saturation; steep slopes will tend to decrease it. The topographic index, a/s, discussed above, reflects these counteracting tendencies. High values of the index (high contributing areas, low slope) will indicate, other things being equal, a higher likelihood of saturated conditions occurring; low values of the index indicate the reverse.

Channel Extension

It is not only runoff source areas on the hillslopes that exhibit dynamic extension during storm periods. Many studies have emphasized the role of extension of the ephemeral channel network in the generation of storm runoff (e.g., Hewlett, 1974). The area of the channel itself and the immediately adjacent *riparian* area can, in some catchments in which the storm hydrograph represents a volume equal to only a small proportion of the incident rainfall (a small *runoff coefficient*), be the most important source area of storm runoff.

4 EFFECT OF HETEROGENEITY

Heterogeneity of Hillslope Forms

Topography is an important control on runoff generation in catchments where downslope flows are an important control on the runoff response. The topographic index provides one very simple approach to identify likely runoff production areas. This will be particularly true for saturation excess runoff production or subsurface stormflows where topographic convergence is important. It may also be true for infiltration excess runoff production where a catenary relationship between soil textural properties and topographic position has developed due to translocation of clays and other long-term processes. The topographic index will be limited as a predictor wherever flow lines in the subsurface depart radically from the surface topography, as in fractured systems or deep groundwater systems, or where the soil is relatively dry such that the effective upslope contributing areas for subsurface flow are very small (e.g., Barling et al., 1994).

Heterogeneity of Soil Characteristics

It has also been noted that variability of soil characteristics may have an important control on runoff production for all the suggested mechanisms. This will be true for the variability of soil series at the catchment scale and also for the variability in characteristics found within a soil series that may not show clear spatial patterns. Variability in soil characteristics associated with the occurrence of vegetation may also be important. Dunne et al. (1991), for example, suggest that the infiltration capacities of the soil beneath the plants of a sparse vegetation canopy can be much higher than between the plants, to the extent that depths of overland flow generated during an event may be greatly affected by local infiltration beneath the plants. Similarly in the Tiger Bush area of the HAPEX-SAHEL experiment in Niger, it is thought that surface runoff from bare soil areas may serve to increase the water available to the adjacent stripes of Tiger Bush (Peugeot et al., 1997).

In all these cases, it is clear that the knowledge of mean soil hydrological characteristics may not be sufficient to understand surface runoff production. Rather the distribution of characteristics within a catchment area may be important. Even with detailed information about soil properties, there is some doubt that the nature of runoff production is truly predictable in detail, especially for infiltration excess type mechanisms. The plot studies of Hjelmfelt and Burwell (1984), for example, suggested that measured surface runoff from adjacent plots may be unpredictable while the study of Loague (1990) and Loague and Kyriakidis (1997) on the R-5 catchment at Chickasha, Ohio, revealed the difficulty of modeling spatially heterogeneous infiltration excess runoff production even on a small catchment with detailed soil measurements.

Heterogeneity of Vegetation

The effects of vegetation on the effective rainfall intensities at ground level and on infiltration rates have already been noted. There are also other effects associated with the heterogeneity of vegetation. If surface runoff is generated, then the effective roughness of the surface may be controlled primarily by the surface vegetation. On surfaces covered by grasses, for example, this may lead to very low velocities for surface runoff (velocities as low as 20 m/h have been measured in the field). The vegetation may also protect the surface from erosion by both rainsplash and flowing water.

Heterogeneity of Precipitation Inputs

Whatever the nature of the surface or subsurface runoff generation processes, the amount of runoff generation will predominantly be determined by the forcing of the rainstorm inputs and the state of antecedent wetness of the catchment. Incident precipitation intensities can vary greatly in space, particularly during convective rainstorm events when patterns of intensities depend strongly on the growth and decay of storm cells and the overall movement of the storm (e.g., Smith et al., 1996).

Runoff generation may depend strongly on patterns of rainfall intensity as suggested, for example, by simulations of the Walnut Gulch experimental catchment in Arizona by Goodrich et al. (1994).

5 IMPORTANCE OF RUNOFF IN GRID-SCALE LAND SURFACE MODELING FOR GCMs

In all past general circulation model (GCM) land surface model components, runoff has not been considered to be very important. It has generally been treated simply as an excess of water that magically disappears from the local water balance. In real catchments, of course, runoff does not disappear but may have an effect on the hydrology and energy balance of areas downslope or downstream. Far more effort, computer time, and parameters have been devoted to formulating the controls on the local energy fluxes of latent and sensible heat than the controls on runoff production. There are several good reasons why runoff production requires more attention, and this lack of attention is now starting to be redressed in the development of so-called macroscale hydrological models

Runoff in many environments is a major part of the water balance and in areas where availability of water is a critical control on latent heat fluxes, then estimating correctly the partitioning of the water balance into that part that is runoff and that part that is available for evapotranspiration may be crucial. This may be a more difficult problem than estimating an areal average evapotranspiration flux since, as is clear from the discussion above, runoff generation has an important spatial dimension due to dependencies on patterns of rainfall inputs, patterns of soil characteristics, and the effects of topography. An important reason why it might be important to improve the runoff generation algorithms in GCMs is that there are long-term discharge measurements available for model evaluation in some catchments at a range of scales and in a wide range of climatic conditions. There is also an increasing recognition that transfers across the grid elements of a GCM, by either surface or regional groundwater flows, may contribute to the controls on patterns of inputs of freshwater to the oceans. At least in areas subject to seasonal flooding (such as the Amazon, Nile, Niger, and other large river basins), such transfers may also control the magnitude of latent heat fluxes over extensive areas.

Macroscale representations of runoff generation are still at an early stage and, given the dependencies on complex spatial heterogeneities and antecedent conditions, it is still not clear as to what an appropriate strategy will be for the formulation and parameter identification of a large-scale model.

REFERENCES

Barling, R. D., I. D. Moore, and R. B. Grayson, A quasi-dynamic wetness index for characterising the spatial distributioin of zones of surface saturation and soil water content, *Water Resour. Res.*, *30*, 1029–1044, 1994.

Bazemore, D. E., K. N. Eshleman, and K. J. Hollenbeek, The role of soil water in stormflow generation in a forested headwater catchment: Synthesis of natural tracer and hydrometeric evidence, *J. Hydrol.*, *162*, 47–75, 1994.

Beven, K. J., TOPMODEL: A critique, *Hydrol. Process.*, *11*(9), 1069–1085, 1997.

Beven, K. J., *Rainfall-Runoff Modelling—The Primer*, Wiley, Chichester, 2001.

Beven, K. J., and P. E. Germann, Macropores and water flow in soils, *Water Resour. Res.*, *18*, 1311–1325, 1982.

Beven, K. J., R. Lamb, P. Quinn, R. Romanowicz, and J. Freer, TOPMODEL, in V.P. Singh (Ed.), *Computer Models of Watershed Hydrology*, Water Resource Publications, Co., 1995, pp. 627–668.

Bonell, M., D. A. Gilmour, D. F. Sinclair, Soil hydraulic properties and their effect on surface and subsurface water transfer in a tropical rainforest catchment, *Hydrol. Sci. Bull.*, *16*, 1–18, 1981.

Dunne, T., Field studies of hillslope flow processes, in M. J. Kirkby (Ed.), *Hillslope Hydrology*, Wiley, Chichester, 1978, pp. 227–293.

Dunne, T., W. Zhang, and B. F. Aubrey, Effects of rainfall, vegetation and microtopography on infiltration and runoff, *Water Resour. Res.*, *27*, 2271–2286, 1991.

Famiglietti, J. S., E. F. Wood, M. Sivapalan, and D. J. Thongs, A catchment scale water balance model for FIFE, *J. Geophys. Res.*, *97*(D17), 18997–19007, 1992.

Genereux, D. P., H. F. Hemond, and P. J. Mulholland, Spatial and temporal variability in streamflow generation on the West Fork of Walker Branch Watershed, *J. Hydrol.*, *142*, 137–166, 1993.

Goodrich, D. C., T. J. Schmugge, T. J. Jackson, C. L. Unkrich, T. O. Keefer, R. Parry, L. B. Bach, and S. A. Amer, Runoff simulation sensitivity to remotely sensed initial soil moisture content, *Water Resour. Res.*, *30*, 1393–1405, 1994.

Hewlett, J. D., Comments on letters relating to "Role of subsurface flow in generating surface runoff. 2. Upstream source areas" by R. Allen Freeze, *Water Resour. Res.*, *10*, 605–607, 1974.

Hjelmfelt, A. T., and R. E. Burwell, Spatial variability of runoff, *J. Irrig. Drain. Div. ASCE*, *110*, 46–54, 1984.

Horton, R. E. The role of infiltration in the hydrological cycle, *Trans. Am. Geophys. Union*, *14*, 446–460, 1933.

Hursh, C. R., Subsurface flow, *Trans. Am. Geophys. Union*, *25*, 743–746, 1944.

Kirkby, M. J., Implications for sediment transport, in M. J. Kirkby (Ed.), *Hillslope Hydrology*, Wiley, Chichester, 1978, pp. 325–363.

Loague, K. M., R-5 revisited: 2. Reevaluation of a quasi-physically based rainfall-runoff model with supplemental information, *Water Resour. Res.*, *26*, 973–987, 1990.

Loague, K. M., and P. C. Kyriakidis, Spatial and temporal variability in the R-5 infiltration data set: Déjà vu and rainfall-runoff simulations, *Water Resour. Res.*, *33*, 2883–2896, 1997.

McDonnell, J. J., A rationale for old water discharge through macropores in a steep humid catchment, *Water Resour. Res.*, *26*(11), 2821–2832, 1990.

O'Loughlin, E. M. Saturation regions in catchments and their relations to soil and topographic properties, *J. Hydrol.*, *53*, 229–246, 1981.

Peugeot, C., M. Esteves, S. Galle, J. L. Rajot, and J. P. Vandervaere, Runoff generation processes: Results and analysis of field data collected at the East Central Supersite of the HAPEX-Sabel experiment, *J. Hydrol.*, *188*, 203–223, 1997.

Rasmussen, T. C., R. H. Baldwin, J. F. Dowd, and A. G. Williams, Tracer versus pressure wave velocities through unsaturated saprolite, *Soil Sci. Soc. Am. J.*, *64*, 75–85, 2000.

Rawls, W. J., and D. L. Brakensiek, Estimation of soil hydraulic properties, in H. J. Morel-Seytoux (Ed.), *Unsaturated Flow in Hydrologic Modeling*, Reidel, Dordrecht, 1989, pp. 275–300.

Römkens, M. J. M., S. N. Prasad, and F. D. Whisler, Surface sealing and infiltration, in M. G. Anderson and T. P. Burt (Eds.), *Processs Studies in Hillslope Hydrology*, 1990, pp. 127–172, Wiley, Chichester.

Sklash, M. G., Environmental isotope studies of storm and snowmelt runoff generation, in M. G. Anderson, and T. P. Burt (Eds.), *Process Studies in Hillslope Hydrology*, 1990, pp. 401–436, Wiley, Chichester.

Smith, J. A., M. L. Baeck, M. Steiner, and A. J. Miller, Catastrophic rainfall from an upslope thunderstorm in the central Appalachians, the Rapidan storm of June 27, 1995, *Water Resour. Res.*, *32*, 3099–3113, 1996.

Weyman, D. R., Throughflow on hillsides and its relation to the stream hydrograph, *Bull. Int. Assoc. Sci. Hydrol.*, *15*, 25–33, 1970.

CHAPTER 30

FLOW ROUTING

D. L. FREAD

1 INTRODUCTION

Flow routing is a mathematical method (model) to predict the changing magnitude, speed, and shape of a flood wave at one or more locations along waterways such as rivers, reservoirs, canals, or estuaries. The flood wave can emanate from precipitation runoff (rainfall or snowmelt), reservoir releases (spillway flows or dam failures), and tides (astronomical and/or wind generated).

Flow routing has long been of vital concern and many ways have been developed to predict the characteristic features of a flood wave to improve the transport of water through natural or man-made waterways and to determine necessary actions to protect life and property from the effects of flooding. Commencing with investigations by Newton (1687), Laplace (1776), Poisson (1816), Boussinesq (1871), and culminating in the one-dimensional equations of unsteady flow derived by Barré de Saint-Venant (1871), the theoretical foundation for flow routing was essentially achieved. The original Saint-Venant equations are the conservation of mass equation:

$$\partial(AV)/\partial x + \partial A/\partial t = 0 \tag{1}$$

and the conservation of momentum equation:

$$\partial V/\partial t + V\ \partial V/\partial x + g(\partial h/\partial x + S_f) = 0 \tag{2}$$

in which t is time, x is distance along the longitudinal axis of the waterway, A is cross-sectional area, V is velocity, g is the gravity acceleration constant, and h is the water-surface elevation above a datum; S_f is the friction slope, which may now be

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

evaluated using a steady flow empirical formula such as the Manning equation (Manning, 1889; Chow, 1959), i.e.,

$$Q = \mu A R^{2/3} S_0^{1/2} / n \tag{3}$$

in which $Q = AV$ is discharge or flow, $R = A/P$ is the hydraulic radius, and P is the wetted perimeter of the cross section, S_0 is the channel bottom slope (dimensionless), μ is a units conversion factor, i.e., 1.49 for U.S. units or 1.0 for SI, and n is the Manning roughness (friction) coefficient. Equations (1) and (2) are quasi-linear hyperbolic partial differential equations with two dependent parameters (V and h) and two independent parameters (x and t); A is a known function of h, and S_f is a known function of V and h. Derivations of the Saint-Venant equations can be found in the following references: Stoker (1957), Henderson (1966), Strelkoff (1969), and Liggett (1975).

Due to the mathematical complexity of the Saint-Venant equations (no analytical solution is known), simplifications were necessary to obtain feasible solutions for the salient characteristics of a propagating flood wave. This approach produced a profusion of simplified flow routing models. The simplified flow routing models may be categorized as: (I) purely empirical, (II) storage routing, based on the conservation of mass and an approximate relation between flow and storage, and (III) hydraulic, i.e., based on the conservation of mass and a simplified form of the conservation of momentum equation (2).

Categories I and II are further classified as lumped flow routing techniques in which the flow is computed as a function of time, only at the most downstream location of routing reaches along the waterway. Category III can be classified as distributed flow routing techniques in which flow and depth or water-surface elevation are computed as a function of time at frequent locations within routing reaches along the waterway. During the last two decades dynamic hydraulic distributed flow routing methods based on numerical solutions of the complete Saint-Venant equations have become economically feasible as a result of advances in computing equipment and improved numerical solution techniques. Following is a brief description of some of the more popular storage routing models as well as both simplified and dynamic hydraulic flow routing models.

2 STORAGE ROUTING MODELS

Significant river improvement projects in the early 1900s provided the impetus for development of an array of simplified flow routing methods. These have been termed storage routing models. They are based on the conservation of mass equation (1) written in the following form:

$$\bar{I} - \bar{O} = \Delta \bar{S} / \Delta t \tag{4}$$

in which $\Delta\bar{S}$ is the change in storage within the routing reach during a Δt time increment, $\bar{I} = 0.5[I(t) + I(t + \Delta t)]$, and $\bar{O} = 0.5[O(t) + O(t + \Delta t)]$; the storage ($\bar{S}$) is assumed to be related to inflow ($\bar{I}$) and/or outflow ($\bar{O}$), i.e.,

$$\bar{S} = \bar{K}[\bar{X}\bar{I} + (1 - \bar{X})\bar{O}] \tag{5}$$

in which $\bar{K}$ is a storage constant with dimensions of time, and $\bar{X}$ is a weighting coefficient, $0 \leq \bar{X} \leq 1$. Storage routing models are limited to typical flood routing applications where the outflow and water-surface elevation relation is essentially single valued, and the waterways are not mild sloping ($S_0 > 0.002$). Thus, backwater effects from tides, significant tributary inflow, and dams or bridges are not considered by these models, nor are they well-suited for rapidly changing unsteady flows such as dam-break flood waves, reservoir power releases, or hurricane storm surges. Generally, storage routing models have two parameters that can be calibrated to effectively reproduce the flood wave speed and its attenuated peak. The calibration requires that most storage routing model applications be limited to where observed inflow–outflow hydrographs exist. When using the observed hydographs to calibrate the routing coefficients, variations in flood wave shapes within the observed data set are not considered, and only the average wave shape is reflected in the fitted routing coefficients.

Reservoir Storage Routing Model

Storage routing models applicable to reservoirs, which have essentially level water-surface profiles, can be developed by assuming $\bar{X}$ to be zero in Eq. (5), i.e., storage is dependent only on outflow. Expressing the term $\Delta\bar{S}/\Delta t$ in Eq. (4) as the product of reservoir surface area (S_a), which is a known function of water-surface elevation (h) and the change of h over a j Δt time step, i.e.,

$$\Delta\bar{S}/\Delta t = 0.5(S_a^j + S_a^{j+1})(h^{j+1} - h^j)/\Delta t^j \tag{6}$$

Now denoting $\bar{O}$ (outflow) as $\bar{Q}$ (discharge), the following reservoir routing model (Fread, 1977) is obtained:

$$0.5(I^j + I^{j+I}) - 0.5(Q^j + Q^{j+1}) - 0.5(S_a^j + S_a^{j+1})(h^{j+1} - h^j)/\Delta t^j = 0 \tag{7}$$

The inflows (I) at times j and $j + 1$ are known from the specified inflow hydrograph; the outflow (Q^j) at time j can be computed from the known water-surface elevation (h^j) and an appropriate spillway discharge equation. The surface area (S_a^j) can be determined from the known value of h^j. The unknowns in the equation consist of h^{j+1}, Q^{j+1}, and S_a^{j+1}; the latter two are known nonlinear functions of h^{j+1}. Hence, Eq. (7) can be solved for h^{j+1} by an iterative method such as Newton–Raphson, i.e.,

$$h_{k+1}^{j+1} = h_k^{j+1} - f(h_k^{j+1})/f'(h_k^{j+1}) \tag{8}$$

in which k is the iteration counter; and $f(h_k^{j+1})$ is the left-hand side of Eq. (7) evaluated with the first estimate for h_k^{j+1}, which for $k = 1$ is either h^j or a linear extrapolated estimate of h^{j+1}; $f'(h_k^{j+1})$; is the derivative of Eq. (7) with respect to h^{j+1}. It can be approximated by using a numerical derivative as follows:

$$f'(h_k^{j+1}) = [f(h_k^{j+1} + \varepsilon) - f(h_k^{j+1} - \varepsilon)]/[(h_k^{j+1} + \varepsilon) - (h_k^{j+1} - \varepsilon)] \tag{9}$$

in which ε is a small value, say 0.1 ft (0.03 m). Using Eq. (8), only one or two iterations are usually required to solve Eq. (7) for h^{j+1}. Initially, the reservoir pool elevation (h^j) must be known to start the computational process. Once h^{j+1} is obtained, Q^{j+1} can be computed from the spillway discharge equation, $Q = f(h_k^{j+1})$.

Level-pool routing is less accurate as the reservoir length increases, as the reservoir mean depth decreases, and as the time of rise of the inflow hydrograph decreases (Fread, 1992). This inaccuracy can have significant economic effects on water control management practices (Sayed and Howard, 1983).

Muskingum Model

A widely used hydrologic flow routing model is the Muskingum model developed by using Eq. (5), with nonzero values for both $\bar{K}$ and $\bar{X}$, for the storage relationship. Substituting this information into Eq. (4), the following is obtained for computing $O(t + \Delta t)$:

$$O(t + \Delta t) = C_1 I(t + \Delta t) + C_2 I(t) + C_3 O(t) \tag{10}$$

where

$$C_0 = \bar{K} - \bar{K}\bar{X} + \Delta t/2 \tag{11}$$

$$C_1 = -(\bar{K}\bar{X} - \Delta t/2)/C_0 \tag{12}$$

$$C_2 = (\bar{K}\bar{X} + \Delta t/2)/C_0 \tag{13}$$

$$C_3 = (\bar{K} - \bar{K}\bar{X} - \Delta t/2)/C_0 \tag{14}$$

and where $C_1 + C_2 + C_3 = 1$ and $\bar{K}/3 \leq \Delta t \leq \bar{K}$ is usually the range for Δt.

Equation (10) is the widely used Muskingum routing model first developed by McCarthy (1938). The parameters $\bar{K}$ and $\bar{X}$ are determined from observed inflow–outflow hydrographs using least squares or its equivalent, the graphical method, or other techniques (Singh and McCann, 1980). Among the many descriptions and variations of the Muskingum model are Chow (1964); Chow et al. (1988), Strupczewski and Kundzewicz (1980), Dooge et al. (1982), and Linsley et al. (1986).

Muskingum–Cunge Model

A significant improvement of the Muskingum model was developed by Cunge (1969) known as the Muskingum–Cunge model. This increased the Muskingum

model's accuracy and made it applicable in situations where observed inflow and outflow hydrographs were not available for calibration and enabled it to be changed from a lumped to a distributed flow routing model. Cunge derived Eq. (10) using the assumption of a single-valued $Q(h)$ relation, the classical kinematic wave equation [see Eq. (25)], and applying a four-point implicit finite-difference approximation technique. Equation (10) is rewritten where the flows $I(t)$, $I(t+\Delta t)$, $O(t)$, and $O(t+\Delta t)$ are replaced by Q_i^j, Q_i^{j+1}, Q_{i+1}^j, and Q_{i+1}^{j+1}, respectively, i.e.,

$$Q_{i+1}^{j+1} = C_1 Q_i^{j+1} + C_2 Q_i^j + C_3 Q_{i+1}^j + C_4 \tag{15}$$

$$C_4 = 0.5[q(t) + q(t+\Delta t)]\Delta x\, \Delta t / C_0 \tag{16}$$

Equation (15) has been expanded to include effects of lateral flow (q) along the ΔX routing reach; and where the following expressions for $\bar{K}$ and $\bar{X}$ are determined:

$$\bar{K} = \Delta x / \bar{c} \tag{17}$$

$$\bar{X} = 0.5[1 - \bar{Q}/(\bar{c}\bar{B}S_e\, \Delta x)] \tag{18}$$

where

$$\bar{c} = dQ/dA \tag{19}$$

in which $\bar{c}$ is the kinematic wave speed, Δx is the reach length, and S_e is the energy slope approximated by evaluating S_0 in Eq. (3) for the initial flow condition. The overbar indicates the variable is averaged over the Δx reach and over the Δt time step. Equation (19) may be expressed in an alternative form, i.e.,

$$\bar{c} = K'\bar{Q}/\bar{A} \tag{20}$$

where

$$K' = \tfrac{5}{3} - \tfrac{2}{3}(d\bar{B}/dy)\bar{A}/(\bar{B})^2 \tag{21}$$

in which A is the cross-sectional area, B is the channel width at the water surface, h is the water-surface elevation of the flow, and the Manning equation is used to relate discharge (Q) and depth or water-surface elevation (h). Depending on the cross-section shape, K' may have values in the range $\frac{4}{3} \le K' \le \frac{5}{3}$; the upper value is associated with either a very wide or rectangular channel. Selection of the appropriate time step Δt in secs is given by:

$$\Delta t \le 3600\, T_r / M \tag{22}$$

where T_r is the time of rise in hours of the inflow hydrograph and M is an integer ($10 \le M \le 20$) whose value depends on the extent of variation in the inflow hydro-

graph. The selection of Δx affects the accuracy of the solution. It is related to Δt and is limited by the following inequality (Jones, 1981):

$$\Delta x \leq 0.5c\,\Delta t[1 + (1 + 1.5Q/(Bc^2S_0\,\Delta t))^{1/2}] \tag{23}$$

While the Muskingum–Cunge model does not require observed inflow–outflow hydrographs to establish the routing coefficients as required in the Muskingum model, best results are obtained if the wave speed (c) is determined from actual flow data. Also, the model is restricted to applications where backwater is not significant and discharge-water elevation rating curves do not have significant loops and discharge hydrographs are not rapidly changing with time such as dam-break floods. Nonetheless, the Muskingum–Cunge model (Miller and Cunge, 1975; Ponce and Yevjevich, 1978) is a highly versatile simplified routing model.

3 SIMPLIFIED HYDRAULIC ROUTING MODELS

Prior to computers, or more recently the feasible economical availability of such computational resources, the inability to obtain feasible numerical solutions to the complete Saint-Venant equations resulted in the development of several simplified distributed hydraulic routing models. They are based on the mass conservation equation (1) and various simplifications of the momentum equation (2).

Kinematic Wave Model

The most simple type of distributed hydraulic routing model is the kinematic wave model. It is based on the following simplified form of the momentum equation (2):

$$S_f - S_0 = 0 \tag{24}$$

in which S_0 is the bottom slope of the channel (waterway) and a component of the term, $\partial h/\partial x = \partial y/\partial x - S_0$, in which $\partial y/\partial x$ is assumed to be zero. This assumes that the momentum of the unsteady flow is the same as that of steady, uniform flow described by the Manning equation or a similar expression in which discharge is a single-valued function of depth, i.e., $\partial Q/\partial A = dQ/dA = c$. Also, since $\partial A/\partial t = (\partial A/\partial Q)(\partial Q/\partial t)$ and $Q = AV$, Eq. (1) can be expanded into the classical kinematic wave equation, i.e.,

$$\partial Q/\partial t + c\,\partial Q/\partial x = 0 \tag{25}$$

in which the kinematic wave velocity or celerity (c) is defined by Eq. (20).

Solutions for the kinematic wave equation (25) can be achieved using the method of characteristics or directly by finite-difference approximation techniques of either explicit or implicit types (Chow et al., 1988; Hydrologic Engr. Ctr., 1981; Linsley et al., 1986). The kinematic wave equation does not theoretically account for hydro-

graph (wave) attenuation. It is only through the numerical error associated with the finite-difference solution that attenuation of the hydrograph peak is achieved. Kinematic wave models are limited to applications where single-value, stage-discharge ratings exist—where there are no loop ratings—and where backwater effects are insignificant. Since, in kinematic wave models, flow disturbances can propagate only in the downstream direction, reverse (negative) flows cannot be predicted. Kinematic wave models are appropriately used as components of hydrologic watershed models for overland flow routing of runoff; they are not recommended for channel routing unless the hydograph is very slow rising, the channel slope is moderate to steep, and hydrograph attenuation is quite small. The range of application (with expected modeling errors less than 5%) for kinematic models, including the Muskingum method, is given by the following:

$$T_r S_0^{1.6}/(q_0^{0.2} n^{1.2}) \geq 0.014 \tag{26}$$

in which T_r is the time (in hours) of rise of the wave (hydrograph), i.e., the interval of time from beginning of significant rise to when the peak occurs; S_0 is the bottom slope (in ft/ft), q_0 is the unit-width discharge (Q/B) (in ft^2/s), and n is the Manning roughness coefficient (Fread, 1985, 1992).

Diffusion Wave Model

Another simplified distributed routing model, known as the diffusion wave (zero inertia) model, is based on Eq. (1) along with an approximation of the momentum equation that retains only the last two terms in Eq. (2), i.e.,

$$\partial h/\partial x + S_f = 0 \tag{27}$$

Finite-difference approximation techniques, both explicit and implicit (Strelkoff and Katopodes, 1977), have been used to obtain simultaneous solutions to Eqs. (1) and (27). The diffusion-simplified routing model considers backwater effects; however, its accuracy is deficient for very fast rising hydographs, such as those resulting from dam failures, hurricane storm surges, or rapid reservoir releases, which propagate through mild to flat sloping waterways with medium to small Manning's n. The range of application (with expected modeling errors less than 5%) for the diffusion models, including the Muskingum–Cunge model, is given by the following (Fread, 1992):

$$T_r S_0^{0.7} n^{0.6}/q_0^{0.4} \geq 0.0003 \tag{28}$$

4 DYNAMIC ROUTING MODEL

When the complete Saint-Venant equations [(1) and (2)] are used, the routing model is known as a dynamic routing model. With the advent of high-speed computers, Stoker (1953) and Isaacson et al. (1954, 1956) first attempted to use the complete Saint-Venant equations for routing Ohio River floods. Since then, much effort has been expended on the development of dynamic routing models. Many models have been reported in the literature (Fread, 1985, 1992; Liggett and Cunge, 1975).

Dynamic routing models can be categorized as characteristic or direct methods of solving the Saint-Venant equations. In the characteristic methods, the Saint-Venant equations are first transformed into an equivalent set of four ordinary differential equations that are then approximated with finite differences to obtain solutions. Characteristic methods (Abbott, 1966; Henderson, 1966; Streeter and Wylie, 1967; Baltzer and Lai, 1968) have not proven advantageous over the direct methods for practical flow routing applications.

Direct methods can be classified further as either explicit or implicit. Explicit schemes (Stoker, 1953, 1957; Isaacson et al., 1954; Garrison et al., 1969; Dronkers, 1969; Strelkoff, 1970; Liggett and Cunge, 1975; Veissman et al., 1977; Linsley et al., 1986) transform the differential equations into a set of algebraic equations that are solved sequentially for the unknown flow properties at each cross section at a given time. However, implicit schemes (Preissman, 1961; Amein and Fang, 1970; Strelkoff, 1970; Fread, 1973, 1977, 1978, 1985; Liggett and Cunge, 1975; Cunge et al., 1980; Schaffranek, 1987; Fread and Lewis, 1998; Chow et al., 1988; Barkow, 1990) transform the Saint-Venant equations into a set of algebraic equations that must be solved simultaneously for all Δx computational reaches at a given time; this set of simultaneous equations may be either linear or nonlinear, the latter requiring an iterative solution procedure.

Explicit methods, although simpler in application, are restricted by numerical stability considerations. Stability problems arise when inevitable errors in computational roundoff and those introduced in approximating the partial differential equations via finite differences accumulate to the point that they destroy the usefulness and integrity of the solution, if not the total breakdown of the computations, by creating artificial oscillations of length about $2\Delta x$ in the solution. Due to stability requirements, explicit methods require very small computational time steps on the order of a few seconds or minutes depending on the ratio of the computational reach length (Δx) to the minimum dynamic wave celerity (u), i.e., $\Delta t \leq \Delta x/u$, where $u = V + (gA/B)^{1/2}$. This is known as the Courant condition, and it restricts the time step to less than that required for an infinitesimal disturbance to travel the Δx distance. Such small time steps cause explicit methods to be inefficient in the use of computer time.

Implicit finite-difference techniques, however, have no restrictions on the size of the time step due to mathematical stability; however, numerical convergence (accuracy) considerations require some limitation in time step size (Fread, 1974; Cunge et al., 1980). Implicit techniques are generally preferred over explicit because of their computational efficiency. Therefore, an implicit scheme will be subsequently

described in detail herein. Rather than using finite-difference approximation techniques to solve the Saint-Venant equations, finite-element techniques (Gray et al., 1977; DeLong, 1986, 1989) can be used; however, their greater complexity offsets any apparent advantages when compared to a weighted, four-point implicit finite-difference scheme (described later) for solving the one-dimensional flow equations. However, finite-element techniques are often applied to two- and three-dimensional flow computations.

Saint-Venant Equations

A modified and expanded form (Fread, 1988, 1992) of the original one-dimensional Saint-Venant equations [(1) and (2)] consist of the conservation of mass equation, i.e.,

$$\partial Q/\partial x + \partial s_c(A + A_0)/\partial t - q = 0 \tag{29}$$

and the momentum equation, i.e.,

$$\sigma[\partial(s_m Q)/\partial t + \partial(\beta Q^2/A)/\partial x] + gA(\partial h/\partial x + S_f + S_{ec} + S_i) + L + W_f B = 0 \tag{30}$$

where Q is discharge, h is the water-surface elevation, A is the active cross-sectional area of flow, A_0 is the inactive (off-channel storage) cross-sectional area, s_c and s_m are area-weighted and conveyance-weighted sinuosity factors, respectively (DeLong, 1986, 1989), which correct for the departure of a sinuous in-bank channel from the x-axis of the floodplain, x is the longitudinal mean-flow-path distance measured along the center of the waterway (channel and floodplain), t is time, q is the lateral inflow or outflow per lineal distance along the waterway (inflow is positive and outflow is negative), σ is a numerical filter ($0 \leq \sigma \leq 1$, usually $\sigma = 1$) to enable the equations to properly handle mixed subcritical/supercritical flows (Fread et al., 1996) during the numerical solution (see the discussion on subcritical/supercritical mixed flow for more on σ later in this chapter), β is the momentum coefficient for nonuniform velocity distribution within the cross section, g is the gravity acceleration constant, S_f is the boundary friction slope, S_{ec} is the expansion/contraction (large eddy loss) slope, and S_i is the viscous dissipation slope for mud/debris flows.

Friction Slope. The boundary friction slope (S_f) is evaluated by rearranging the Manning Eq. (3) for uniform, steady flow into the following form:

$$S_f = n^2|Q|Q/(\mu^2 A^2 R^{4/3}) = |Q|Q/K^2 \tag{31}$$

in which n is the Manning coefficient of frictional resistance (Chow, 1959; Barnes, 1967; Arcement and Schneider, 1984; Jarrett, 1984; and Fread, 1989), R is the hydraulic radius, μ is a units conversion factor (1.49 for U.S. units and 1.0 for SI), and K is the channel conveyance factor. The absolute value of Q is used to correctly account for the possible occurrence of reverse (negative) flows. The

conveyance formulation is preferred (for numerical and accuracy considerations) for composite channels having wide, flat overbanks or floodplains in which K represents the sum of the conveyance of the channel (which is corrected for sinuosity effects by dividing by s_m), and the conveyances of left and right floodplain areas.

When the conveyance factor (K) is used to evaluate S_f, the river channel/valley cross-sectional properties are designated as left floodplain, channel, and right floodplain rather than as a composite channel/valley section. Special orientation for designating left or right is not required as long as consistency is maintained. The conveyance factor is evaluated as follows (Fread and Lewis, 1998):

$$K = K_l + K_c + K_r \tag{32}$$

where:

$$K_l = \frac{\mu}{n_l} A_l R_l^{2/3} \tag{33}$$

$$K_c = \frac{\mu A_c R_c^{2/3}}{n_c s_m^{1/2}} \tag{34}$$

$$K_r = \frac{\mu}{n_r} A_r R_r^{2/3} \tag{35}$$

in which the subscripts l, c, and r designate left floodplain, channel, and right floodplain, respectively.

Sinuosity Factors. The area-weighted and conveyance-weighted sinuosity factors (s_c and s_m, respectively) in Eqs. (29), (30), and (34) represent the ratio(s) of the flow-path distance along a meandering channel to the mean-flow-path distance along the floodplain. Initially, only one sinuosity factor (s_k) is specified as varying only with each Jth depth of flow ($J = 1, 2, \ldots, \hat{J}$, where $\hat{J}$ is the number of user-specified tabular top widths (B) versus h values, which describe the cross-section geometry), but then this is recomputed within the model according to the following relations:

$$s_{cJ} = \sum_{k=2}^{k=J} (\Delta A_{lk} + \Delta A_{ck} s_k + \Delta A_{rk}) / (A_{lJ} + A_{cJ} + A_{rJ}) \tag{36}$$

$$s_{mJ} = \sum_{k=2}^{k=J} (\Delta K_{lk} + \Delta K_{ck} s_k + \Delta K_{rk}) / (K_{lJ} + K_{cJ} + K_{rJ}) \tag{37}$$

in which $\Delta A = A_{J+1} - A_J$, and s_k represents the sinuosity factor for a differential portion of the flow between the Jth depth and the $J + 1$th depth, and K is the conveyance factor.

Expansion/Contraction Effects. The term S_{ec} is computed as follows:

$$S_{ec} = k_{ec}\Delta(Q/A)^2/(2g\ \Delta x) \tag{38}$$

in which k_{ec} is the expansion/contraction coefficient (negative for expansion/positive for contraction), which varies from $-1.0/0.4$ for an abrupt change in section geometry to $-0.3/0.1$ for a very gradual, curvilinear transition between cross sections. The Δ represents the difference in the term $(Q/A)^2$ at two adjacent cross sections separated by a distance Δx. If the flow direction changes from downstream to upstream, k_{ec} can be automatically changed (Fread, 1988).

Large floods such as dam-break-generated floods usually have much greater velocities; it is important, especially for nonuniform channels (Rajar, 1978) to include in the Saint-Venant momentum equation (30) the expansion/contraction losses via the S_{ec} term defined by equation (38). The ratio of expansion/contraction action losses (form losses) to the friction losses can be in the range of $0.01 < S_{ec}/S_f < 1.0$. The larger ratios occur for very irregular channels with relatively small n values and for flows with large velocities (dam-break floods).

Momentum Correction Coefficient. The momentum correction coefficient (β) for nonuniform velocity distribution across the cross section is (Chow, 1959)

$$\beta = (K_l^2/A_l + K_c^2/A_c + K_r^2/A_r)/[(K_l + K_c + K_r)^2/(A_l + A_c + A_r)] \tag{39}$$

in which K is conveyance, A is wetted area, and the subscripts l, c, and r denote left floodplain, channel, and right floodplain, respectively. When floodplain properties are not separately specified and the total cross section is treated as a composite section, β can be approximated as $1.0 \leq \beta \leq 1.06$ in lieu of Eq. (39). Also, in this case, S_c and S_m are set to unity in lieu of Eqs. (36) and (37).

Lateral Flow Momentum. The term L in Eq. (30) is the momentum effect of lateral flows and has the following form (Strelkoff, 1969): (a) lateral inflow, $L = -qv_x$, where v_x is the velocity of lateral inflow in the x direction of the main channel flow; (b) seepage lateral outflow, $L = -0.5Q/A$; and (c) bulk lateral outflow, $L = -qQ/A$.

Mud or Debris Flows. The friction loss term (S_i) is included (Fread, 1988) in the momentum equation (30) in addition to S_f to account for viscous dissipation effects of non-Newtonian flows such as mud or debris flows. Also, mine tailings dams, where the viscous contents retained by the dam have non-Newtonian properties, are dam-breach flood applications requiring the use of S_i in Eq. (30). This effect becomes significant only when the solids concentration of the flow is greater than about 40% by volume. For concentrations of solids greater than about 50%, the flow behaves more as a landslide and is not governed by the Saint-Venant equations. S_i is

evaluated for any non-Newtonian flow as follows (Jin and Fread, 1997):

$$S_i = \frac{\tau_y}{\gamma D}\left[1 + \left(\frac{(b+1)(b+2)Q}{(0.74+0.66b)(\tau_y/\kappa)^b DA}\right)^{1/b+0.15}\right] \tag{40}$$

in which γ is the fluid's unit weight, τ_0 is the fluid's yield strength, D is the hydraulic depth (A/B), $b = 1/m$ where m is the exponent of the power function that fits the fluid's stress(τ_s)–strain(dv/dy) properties, and κ is the apparent viscosity or scale factor of the power function, i.e., $\tau_s = \tau_0 + \kappa(dv/dy)^m$. The viscous properties, τ_0 and κ, can be estimated from the solids concentration ratio of the mud flow (O'Brien and Julien, 1984).

Wind Effects. The last term ($W_f B$) in Eq. (30) represents the resistance effect of wind on the water surface (Fread, 1985, 1992); B is the wetted topwidth of the active flow portion of the cross section; and $W_f = V_r|V_r|c_w$, where the wind velocity relative to the water is $V_r = V_w \cos w + V$, V_w is the velocity of the wind, *positive* (+) if opposing the flow velocity and negative (−) if aiding the flow, w is the acute angle the wind direction makes with the x-axis, V is the velocity of the unsteady flow, and c_w is a wind friction coefficient ($1 \times 10^{-6} \leq c_w \leq 3 \times 10^{-6}$). This modeling capability can be used to simulate the effect of potential dam overtopping due to wind setup within a reservoir by applying the Saint-Venant equations to the unsteady flow in a reservoir.

Implicit Four-Point, Finite-Difference Approximations

The extended Saint-Venant equations [(29) and (30)] constitute a system of partial differential equations with two independent variables, x and t, and two dependent variables, h and Q; the remaining terms are either functions of x, t, h, and/or Q, or they are constants. The partial differential equations can be solved numerically by approximating each with a finite-difference algebraic equation; then the system of algebraic equations are solved in conformance with prescribed initial and boundary conditions.

Of various implicit, finite-difference solution schemes that have been developed, a four-point scheme first suggested by Issacson et al. (1954, 1956) and first used by Preissmann (1961) and later by Amein and Fang (1970) and then a weighted version by others (Fread, 1974, 1977, 1985, 1988; Cunge et al., 1980) is most advantageous. It is readily used with unequal distance steps, its stability–convergence properties are conveniently modified, and boundary conditions are easily applied.

Space–Time Plane. In the weighted four-point implicit scheme, the continuous x–t region in which solutions of h and Q are sought is represented by a rectangular grid of discrete points as shown in Figure 1. An x–t plane (solution domain) is a convenient means for expressing relationships among the variables. The grid points are determined by the intersection of lines drawn parallel to the x and t axes. Those

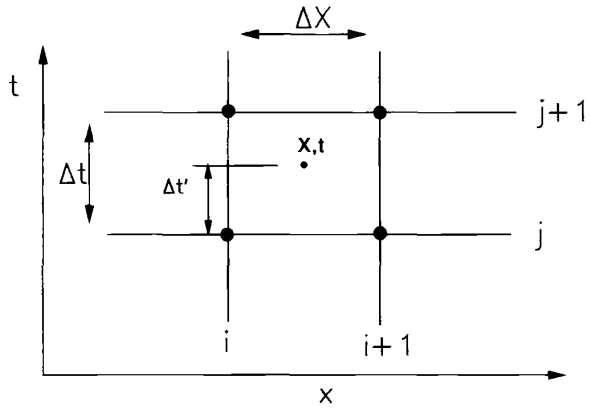


Figure 1 The x–t solution domain for the weighted four-point implicit scheme. See ftp site for color image.

parallel to the t axis represent locations of cross sections; they have a spacing of Δx, which need not be the same between each pair of cross sections. Those parallel to the x axis represent time lines; they have a spacing of Δt, which also need not be the same between successive time lines. Each point in the rectangular network can be identified by a subscript (i), which designates the x position or cross section, and a superscript (j), which designates the particular time line.

Numerical Approximations. The time derivatives are approximated by a forward-difference quotient at point (x', t') centered between the i and $i+1$ lines along the x axis as shown in Figure 1, i.e.,

$$\partial\phi/\partial t = (\phi_i^{j+1} + \phi_{i+1}^{j+1} - \phi_i^j - \phi_{i+1}^j)/2\Delta t_j \tag{41}$$

where ϕ represents any dependent variable or functional quantity (Q, s_c, s_m, A, A_0, q, h). Spatial derivatives are approximated at point (x', t') by a forward-difference quotient located between two adjacent time lines according to weighting factors of θ (the ratio $\Delta t'/\Delta t$ shown in Fig. 1) and $1-\theta$, i.e.,

$$\partial\phi/\partial x = \theta(\phi_{i+1}^{j+1} - \phi_i^{j+1})/\Delta x_i + (1-\theta)(\phi_{i+1}^j - \phi_i^j)/\Delta x_i \tag{42}$$

Nonderivative terms are approximated with weighting factors at the same time level [point (x', t)] where the spatial derivatives are evaluated, i.e.,

$$\phi = \theta(\phi_i^{j+1} + \phi_{i+1}^{j+1})/2 + (1-\theta)(\phi_i^j + \phi_{i+1}^j)/2 \tag{43}$$

The weighted four-point implicit scheme is unconditionally, linearly stable for $\theta \geq 0.5$ (Fread, 1974); however, the sizes of the Δt and Δx computational steps

are limited by the accuracy of the assumed linear variations of functions between the grid points in the x–t solution domain. Values of θ greater than 0.5 dampen parasitic oscillations that have wavelengths of about $2\Delta x$ that can grow enough to invalidate or destroy the solution. The θ weighting factor causes some loss of accuracy as it departs from 0.5, a box scheme, and approaches 1.0, a fully implicit scheme. This effect becomes more pronounced as the magnitude of the ratio $(T_r/\Delta t)$ decreases where T_r is the time of rise of the hydrograph (time interval from beginning of significant rise to peak of the hydrograph). Usually, a θ weighting factor of 0.60 is used to minimize the loss of accuracy while avoiding the possibility of weak (pseudo) instability for θ values of 0.5 when frictional effects are minimal.

Selection of Δt and Δx Computational Parameters. The computational time step (Δt) can be either specified or automatically determined to best suit the most rapidly rising hydrograph occurring within a system of rivers that may contain one or more breaching dams or other dynamic internal boundary conditions. The time step is selected according to the following:

$$\Delta t = T_r/M \tag{44}$$

where T_r is the minimum time of rise (seconds) of any hydrograph that has been specified at upstream boundaries or in the process of being generated at a breaching dam; M is user specified according to the following guidance (Fread, 1993):

$$M = 2.67[1 + \mu' n^{0.9}/(q^{0.1} S_0^{0.45})] \tag{45}$$

in which $\mu' = 3.97$ (3.13 SI units), n is the Manning friction coefficient, q is the peak flow per unit channel width (Q/B), and S_0 is the channel bottom slope; M usually varies within the range, $6 \leq M \leq 40$, with M often assumed to be approximately 20.

The Δx computational distance step can be specified or automatically determined according to the smaller of two criteria (Fread, 1993). The first criterion is

$$\Delta x \leq cT_r/20 \tag{46}$$

in which c is the bulk wave celerity (the celerity or velocity associated with an essential characteristic of the unsteady flow such as the peak of the hydrograph). In most applications, the wave velocity is well approximated as a kinematic wave, and c is estimated as $3/2V$ (V is the flow velocity) or c can be obtained by dividing the distance between two points along the channel by the difference in the times of occurrence of the peak of an observed or computed flow hydrograph at each point. Since c can vary along the channel, and depending upon the extent of this variation, Δx may not be constant along the channel.

The second criterion for selecting Δx is the restriction imposed by rapidly varying cross-sectional changes along the x axis of the waterway. Such expansion/contraction is limited to the following inequality (Samuels, 1985):

$$0.635 < A_{i+1}/A_i < 1.576 \tag{47}$$

This condition results in the following approximation (Fread, 1988) for the maximum computational distance step:

$$\Delta x \leq L'/N' \tag{48}$$

where

$$N' = 1 + 2|A_i - A_{i+1}|/\hat{A} \tag{49}$$

in which L' is the distance between two adjacent (i and $i + 1$) cross sections differing from one another by approximately 50% or greater, $\hat{A}$ is the active cross-sectional area, $\hat{A} = A_{i+1}$ if $A_i > A_{i+1}$ (contracting reach) or $\hat{A} = A_i$ if $A_i < A_{i+1}$ (expanding reach), and N' is rounded to the nearest integer value.

Significant changes in the bottom slope of the waterway also require small distance steps in the vicinity of the change. This is required particularly when the flow changes from subcritical to supercritical or conversely along the waterway. Such changes can require computational distance steps in the range of 50 to 200 ft (15 to 63 m).

Automatic Interpolation. It is convenient to automatically provide linearly interpolated cross sections at a user-specified spatial resolution to increase the spatial frequency at which solutions to the Saint-Venant equations are obtained. This is often required for purposes of attaining numerical accuracy/stability when (a) routing very sharp peaked hydrographs such as those generated by breached dams, (b) when adjacent cross sections either expand or contract by more than about 50%, and (c) where mixed flow changes from subcritical to supercritical or vice versa.

Algebraic Routing Equations. Using the finite-difference operators of Eqs. (41) to (43) to replace the derivatives and other variables in Eqs. (29) and (30), the

following weighted four-point, implicit, nonlinear, finite-difference algebraic equations are obtained:

$$\theta\left[\frac{Q_{i+1}^{j+1}-Q_i^{j+1}}{\Delta x_i}\right]-\theta q_i^{j+1}+(1-\theta)\left[\frac{Q_{i+1}^{j}-Q_i^{j}}{\Delta x_i}\right]-(1-\theta)q_i^{j}$$
$$+\left[\frac{s_{c_i}^{j+1}(A+A_0)_i^{j+1}+s_{c_i}^{j+1}(A+A_0)_{i+1}^{j+1}-s_{c_i}^{j}(A+A_0)_i^{j}-s_{c_i}^{j}(A+A_0)_{i+1}^{j}}{2\Delta t_j}\right]=0 \quad (50)$$

$$\sigma\left[\frac{(s_{m_i}Q_i)^{j+1}+(s_{m_i}Q_{i+1})^{j+1}-(s_{m_i}Q_i)^{j}-(s_{m_i}Q_{i+1})^{j}}{2\Delta t_j}\right]$$
$$+\theta\left[\sigma\left(\frac{(\beta Q^2/A)_{i+1}^{j+1}-(\beta Q^2/A)_i^{j+1}}{\Delta x_i}\right)\right.$$
$$\left.+g\bar{A}_i^{j+1}\left(\frac{h_{i+1}^{j+1}-h_i^{j+1}}{\Delta x_i}+\bar{S}_{f_i}^{j+1}+S_{ec_i}^{j+1}+S_{i_i}^{j+1}\right)+L_i^{j+1}+(W_f\bar{B})_i^{j+1}\right]+(1-\theta) \quad (51)$$

$$\left[\sigma\left(\frac{(\beta Q^2/A)_{i+1}^{j}-(\beta Q^2/A)_i^{j}}{\Delta x_i}\right)+g\bar{A}_i^{j}\left(\frac{h_{i+1}^{j}-h_i^{j}}{\Delta x_i}+\bar{S}_{f_i}^{j}+S_{ec_i}^{j}+S_{i_i}^{j}\right)\right.$$
$$\left.+L_i^{j}+(W_f\bar{B})_i^{j}\right]=0$$

where

$$\bar{A}_i=(A_i+A_{i+1})/2 \quad (52)$$
$$\bar{S}_{f_i}=n^2\bar{\theta}_i|\theta_i|/(\mu^2A_i^2\bar{R}_i^{4/3})=\bar{\theta}_i|\theta_i|/\bar{K}_i \quad (53)$$
$$\bar{Q}_i=(Q_i+Q_{i+1})/2 \quad (54)$$
$$\bar{R}_i=\bar{A}_i/\bar{B}_i \quad (55)$$
$$\bar{B}_i=(B_i+B_{i+1})/2 \quad (56)$$
$$\bar{K}_i=(K_i+K_{i+1})/2 \quad (57)$$

The terms L and W_fB are defined in Eq. (30); terms associated with the jth time line are known from initial conditions or previous time-step computations; and μ in Eq. (53) is defined in Eq. (31). The Δx distance between cross sections is measured along the peak flow path through the waterway.

Solution Procedure

The flow equations are expressed in finite-difference form for all Δx_i reaches between the first and last (Nth) cross section ($i = 1, 2, \ldots, N$) along the channel/floodplain and then solved simultaneously for the unknowns (Q and h) at each cross section. In essence, the solution technique determines the unknown quantities (Q and h at all specified cross sections along the waterway) at various times into the future; the solution is advanced from one time to a future time over a finite time interval (time step) of magnitude Δt. Thus, applying Eqs. (50) and (51) recursively to each of the $(N - 1)$ rectangular grids in Figure 1 between the upstream and downstream boundaries, a total of $(2N - 2)$ equations with $2N$ unknowns are formulated. Then, prescribed boundary conditions for subcritical flow [Froude number less than unity, i.e., $\text{Fr} = Q/(A\sqrt{gD}) < 1$], one at the upstream boundary and one at the downstream boundary, provide the two additional and necessary equations required for the system to be determinate. Since disturbances can propagate only in the downstream direction in supercritical flow ($\text{Fr} > 1$), two upstream boundary conditions and no downstream boundary condition are required for the system to be determinate when the flow is supercritical throughout the routing reach. The boundary conditions are described later. Due to the nonlinearity of Eqs. (50) and (51) with respect to Q and h, an iterative, highly efficient quadratic solution technique such as the Newton–Raphson method is frequently used. Other solution techniques linearize Eqs. (50) and (51) via a Taylor series expansion or other means. Convergence of the iterative technique is attained when the difference between successive solutions for each unknown is less than a relatively small prescribed tolerance. Convergence for each unknown at all cross sections is usually attained within about one to five iterations with the majority of solutions obtained within two iterations. A more complete description of the solution method may be found elsewhere (Fread, 1985).

The solution of $2N \times 2N$ simultaneous equations requires an efficient matrix technique for the implicit method to be feasible. One such procedure requiring $38N$ computational operations $(+, -, *, /)$ is a compact, penta-diagonal Gaussian elimination method (Fread, 1971, 1985) that makes use of the banded structure of the coefficient matrix of the system of equations. This is essentially the same as the double sweep elimination method (Liggett and Cunge, 1975; Cunge et al., 1980).

When flow is everywhere and at all times supercritical, the solution technique previously described can be somewhat simplified. Two boundary conditions are required at the upstream boundary and none at the downstream boundary since flow disturbances cannot propagate upstream in supercritical flow. The unknown h and Q at the most upstream cross section are determined from the two boundary equations. Then, cascading from upstream to downstream, Eqs. (50) and (51) are solved for the two unknowns (h_{i+1} and Q_{i+1}) at each cross section by using Newton–Raphson iteration applied recursively to the two nonlinear equations, with $\sigma = 0$ in Eq. (51).

Initial Conditions

Values of water-surface elevation (h) and discharge (Q) for each cross section must be specified initially at time $t = 0$ to obtain solutions to the Saint-Venant equations. Initial conditions may be obtained from any of the following: (a) observations at gaging stations and using interpolated values between gaging stations for intermediate cross sections in large rivers; (b) computed values from a previous unsteady-flow solution (used in real-time flood forecasting); and (c) computed values from a steady-flow backwater solution. The backwater method is most commonly used in which the steady discharge at each cross section is determined by:

$$Q_{i+1} = Q_i + q_i \Delta x_i \qquad i = 1, 2, 3, \ldots, N-1 \tag{58}$$

in which Q_1 is the assumed steady flow at the upstream boundary at time $t = 0$, and q_i is the known average lateral inflow or outflow along each Δx reach at $t = 0$. The water-surface elevations (h_i) are computed according to the following steady-flow simplification of the momentum equation (30):

$$(Q^2/A)_{i+1} - (Q^2/A)_i + g\bar{A}_i(h_{i+1} - h_i + \Delta x_i \bar{S}_{f_i}) = 0 \tag{59}$$

in which $\bar{A}$ and $\bar{S}_f$ are defined by Eqs. (52) and (53), respectively. The computations proceed in the upstream direction ($i = N-1, \ldots 3, 2, 1$) for subcritical flow (they must proceed in the downstream direction for supercritical flow). The starting water-surface elevation (h_N) can be specified or obtained from the appropriate downstream boundary condition for the discharge (Q_N) obtained via Eq. (58). The Newton–Raphson iterative solution method (Fread and Harbaugh, 1971) for a single equation and/or a simple, less efficient, but more stable bisection iterative technique can be applied to Eq. (59) to obtain h_i. Due to friction, small errors in the initial conditions will dampen-out after several computational time steps during the solution of the Saint-Venant equations.

Upstream Boundary

Values for the unknowns at external boundaries (the upstream and downstream extremities of the routing reach) of the channel/floodplain, must be specified to obtain solutions to the Saint-Venant equations. In fact, in most unsteady-flow applications, the unsteady disturbance is introduced at one or both of the external boundaries.

A specified discharge time series (hydrograph) of inflow to the most upstream cross section is used as the upstream boundary condition. The hydrograph should not be affected by downstream flow conditions. This hydrograph may be obtained from the following: (1) historical observations, (2) assumed design hydrograph, or (3) a runoff hydrograph from specified rainfall–runoff model using calibrated or

estimated model parameters. The upstream boundary is expressed mathematically as follows:

$$Q_1^{j+1} - Q(t) = 0 \tag{60}$$

in which $Q(t)$ is the specified discharge time series and the subscript indicates the discharge at the first cross section, i.e., the upstream boundary. Equation (60) is used for the upstream boundary if dynamic routing (based on the discretized Saint-Venant equations) commences at this location. However, if the most upstream cross section represents the inlet to an upstream reservoir, a simple routing procedure (reservoir level-pool routing) can be used rather than the considerably more complex dynamic routing if (1) the reservoir is not excessively long and (2) the inflow hydrograph $Q(t)$ is not rapidly changing with time. Level-pool routing errors (Fread, 1992), with a magnitude of less than about 5%, can usually be tolerated.

Downstream Boundary

For subcritical flow, a specified discharge or water-surface elevation time series or a tabular relation between discharge and water-surface elevation (single-valued rating curve) can be used as the downstream boundary condition.

Another downstream boundary condition can be a computed loop-rating curve based on the Manning equation, i.e.,

$$Q_N^{j+1} - \mu/nA_N^{j+1}(R_N^{j+1})^{2/3}(S_{f_N}^j)^{1/2} = 0 \tag{61}$$

The loop is produced by using the friction slope (S_f) rather than the channel bottom slope (S_0) in the Manning equation. The friction slope exceeds the bottom slope during the rising limb of the hydrograph while the reverse is true for the recession limb. The friction slope (S_f) is approximated by using Eq. (30) where L and W_f are assumed to be zero while s_m and β are assumed to be unity (Fread, 1985, 1988, 1992), i.e.,

$$S_{f_N}^j = (Q_N^j - Q_N^{j-1})/(gA_N^j\Delta t^j) - [(Q^2/A)_N^j - (Q^2/A)_{N-1}^j]/(gA_N^j\Delta x_{N-1}) - (h_N^j - h_{N-1}^j)/\Delta x_{N-1} \tag{62}$$

The loop-rating boundary equation allows the unsteady wave to pass the downstream boundary with minimal disturbance by the boundary itself, which is desirable when the routing is terminated at an arbitrary location along the channel/floodplain and not at a location of actual flow control such as a dam or waterfall, or where the flow is affected by downstream backwater conditions produced by tidal action, reservoirs, or tributary inflow.

When the downstream boundary is a stage/discharge relation (rating curve), the flow at the boundary should not be otherwise affected by flow conditions farther downstream. Although there are often some minor effects due to the presence of

cross-sectional irregularities downstream of the chosen boundary location, these usually can be neglected unless the irregularity is so pronounced as to cause significant backwater or drawdown effects. Reservoirs, major tributaries, or tidal effects located below the downstream boundary, which cause backwater effects at the boundary, should be avoided. When either of these situations is unavoidable, the routing reach should be extended downstream to the dam in the case of the reservoir or to a location downstream of where the major tributary enters. Sometimes the routing reach may be shortened by moving the downstream boundary to a location farther upstream where backwater effects are negligible.

Internal Boundaries

Often along the channel/floodplain, there are locations such as a dam, bridge, or waterfall (short rapids) where the flow is rapidly varied in space rather than gradually varied. At such locations (internal boundaries), the Saint-Venant equations are not applicable since gradually varied flow is a necessary condition for their derivation. Empirical water elevation-discharge relations such as weir flow are utilized for simulating rapidly varying flow. At internal boundaries, cross sections are specified for the upstream and downstream extremities of the section where rapidly varying flow occurs. The Δx reach containing an internal boundary requires two internal boundary equations since, as with any other Δx reach, two equations equivalent to the Saint-Venant equations are required. One of the required internal boundary equations represents conservation of mass with negligible time-dependent storage, i.e.,

$$Q_i^{j+1} - Q_{i+1}^{j+1} = 0 \tag{63}$$

Dam. The second equation is usually an empirical, rapidly varied flow relation. If the internal boundary represents a dam, the following equation can be used:

$$Q_i^{j+1} - (Q_s + Q_b)^{j+1} = 0 \tag{64}$$

in which Q_s and Q_b are the spillway and dam-breach flow, respectively. In this way, the flows Q_i and Q_{i+1} and the elevations h_i and h_{i+1} are in balance with the other flows and elevations occurring simultaneously throughout the entire flow system, which may consist of additional downstream dams that are treated as additional internal boundary conditions via Eqs. (63) and (64). In fact, this approach can be used to simulate the progression of a dam-break flood through an unlimited number of reservoirs located sequentially along the valley. The downstream dams may also breach if they are sufficiently overtopped. The spillway flow (Q_s) is computed from the following expression:

$$Q_s = c_s L_s (h_i - h_s)^{1.5} + c_g A_g (h_i - h_g)^{0.5} + c_d L_d (h_i - h_d)^{1.5} + Q_t \tag{65}$$

in which c_s is the uncontrolled spillway discharge coefficient, h_s is the uncontrolled spillway crest, c_g is the gated spillway discharge coefficient, h_g is the centerline elevation of the gated spillway, c_d is the discharge coefficient for flow over the crest of the dam, L_s is the spillway length, and Q_t is a constant outflow term that is head independent or it may be a specified discharge time series. The uncontrolled spillway flow or the gated spillway flow can also be represented as a table of head-discharge values. The gate flow may also be specified as a function of time via a known time series for $A_g(t)$. The breach outflow (Q_b) is computed as broad-crested weir flow (Fread, 1977, 1985, 1988, 1992; Fread and Lewis, 1998), i.e.,

$$Q_b = c_v k_s [3.1 b_i (h_i - h_b)^{1.5} + 2.45 z (h_i - h_b)^{2.5}] \tag{66}$$

in which c_v is a small correction for velocity of approach, b_i is the instantaneous breach bottom width, h_i is the elevation of the water surface just upstream of the structure, h_b is the elevation of the breach bottom in which h_b is assumed to be a function of time (t_b) from beginning of the breach formation time (τ), z is the side slope of the breach, and k_s is the submergence correction factor due to the downstream tailwater elevation (h_t), i.e.,

$$k_s = 1.0 \qquad h^* \leq 0.67 \tag{6.7}$$

$$k_s = 1.0 - 27.8(h^* - 0.67)^3 \qquad h^* > 0.67 \tag{68}$$

where

$$h^* = (h_t - h_b)/(h_i - h_b) \tag{69}$$

Using a parametric description of the breach, the instantaneous breach bottom width (b_i) starts at a point at the crest of the dam and enlarges at a linear or nonlinear rate over the failure time (τ) until the terminal bottom width (b) is attained and the breach bottom has eroded to the minimum elevation, h_{bm}. The instantaneous bottom elevation of the breach (h_b) is described as a function of time (t_b) according to the following:

$$h_b = h_d - (h_d - h_{\text{bm}})(t_b/\tau)^{\rho} \qquad 0 \leq t_b \leq \tau \tag{70}$$

in which h_d is the elevation of the top of the dam, h_{bm} is the final elevation of the breach bottom, which is usually, but not necessarily, the bottom of the reservoir or outlet channel bottom, t_b is the time since beginning of breach formation, and ρ is the parameter specifying the degree of nonlinearity, e.g., $\rho = 1$ is a linear formation rate, while $\rho = 2$ is a nonlinear quadratic rate; the range for ρ is $1 \leq \rho \leq 4$, with the linear rate usually assumed. The interval of time (τ) required for the breach to form is given by $\tau = 0.3 V_r^{0.53}/H_d^{0.9}$ in which $H_d = h_b - h_{\text{bm}}$, V_r is the reservoir volume (acre-ft) from empirical data by Froehlich (1987); the standard error of estimate for τ

is ±0.9h or ±74% of τ (Fread, 1988, 1995). The instantaneous bottom width (b_i) of the breach is given by the following:

$$b_i = b(t_b/\tau)^{\rho} \qquad 0 \le t_b \le \tau \tag{71}$$

in which b is the final width of the breach bottom given by $b = \bar{b} - zH_d$ and $\bar{b} = 9.5k_0(V_rH_d)^{0.25}$ from empirical data by Froehlich (1987) in which $k_0 = 0.7$ for piping and $k_0 = 1.0$ for overtopping; the standard error of estimate for $\bar{b}$ is ±82 ft or ±56% of $\bar{b}$ (Fread, 1988, 1995).

When simulating a dam failure, the actual breach formation can commence when the reservoir water-surface elevation (h) exceeds a user-specified value, h_f. This feature permits the simulation of an overtopping of a dam in which the breach does not form until a sufficient amount of water has passed over the crest of the dam to have eroded away the downstream face of the dam.

If the breach is formed by piping, Eq. (66) is replaced by an orifice equation:

$$Q_b = 4.8A_p(h_i - h_p)^{1/2} \tag{72}$$

where

$$A_p = [b_i + z(h_p - h_b)](h_p - h_b) \tag{73}$$

in which h_p is the specified centerline elevation of the pipe. Each of the terms in Eq. (65) except Q_t may be modified by a submergence correction factor similar to k_s that can be computed by Eqs. (67) to (69), but in Eq. (69) h_b is replaced by h_s, h_g, and h_d, respectively.

Bridge. If the internal boundary represents highway/railway bridges together with their earthen embankments that cross the floodplain, Eqs. (63) and (64) can still be used although Q_s in Eq. (65) is computed by the following contracted bridge flow expression:

$$Q_s = C_b\sqrt{g}A_{i+1}(h_i - h_{i+1})^{0.5} + C_dk_s(h_i - h_c)^{1.5} \tag{74}$$

in which C_b is a coefficient of bridge flow (Chow, 1959), C_d is the coefficient of flow over the crest of the road embankment, h_c is the crest elevation of the embankment, and k_s is similar to Eqs. (67) to (69) except h_b is replaced by h_c. A breach of the embankment is treated the same as with dams.

Levee Overtopping/Floodplain Interactions

Flows that overtop levees located along either or both sides of a main-stem river and/or its tributaries can be treated as lateral flow (q) in Eqs. (29) and (30) where the lateral flow diverted over the levee is computed as broad-crested weir flow. This overtopping flow is corrected for submergence effects if the floodplain water-surface

elevation sufficiently exceeds the levee crest elevation. After the flood peak passes, the overtopping flow may reverse its direction when the floodplain water-surface elevation exceeds the river water-surface elevation, thus allowing flow to return to the river. The overtopping broad-crested weir flow is computed according to the following:

$$q = -c_l k_s (h - h_c)^{3/2} \tag{75}$$

where k_s, the submergence correction factor, is computed as in Eqs. (67) to (69) except $h^* = (h_{fp} - h_c)/(h - h_c)$, in which c_l is the weir discharge coefficient, h_c is the levee-crest elevation, h is the water-surface elevation of the river, and h_{fp} is the water-surface elevation of the floodplain. Flow in the floodplain can affect overtopping flows via the submergence correction factor. Flow may also pass from the waterway to the floodplain through a time-dependent crevasse (breach) in the levee via a breach-flow equation similar to Eq. (66). The floodplain, which is separated from the principal routing channel (river) by the levee, may be treated as (a) a dead-storage area (A_0) in the Saint-Venant equations, in which case Eq. (75) is not relevant, (b) a tributary that receives its inflow as lateral flows (the flows from the river that overtop the levee crest), which are simultaneously dynamically routed along the floodplain, and (c) the flows and water-surface elevations can be computed by using a level-pool routing method particularly if the floodplain is divided into compartments by levees (dikes) or elevated roadways located somewhat perpendicular to the river levee(s).

Supercritical/Subcritical Mixed Flows

Flow can change with either time or distance along the routing reach from supercritical to subcritical while passing through critical flow, or conversely. This "mixed flow" requires special treatment to prevent numerical instabilities in the solution of the Saint-Venant equations. Such a treatment for mixed flows (Fread et al., 1996) is to provide a "local partial inertia" filter, i.e.,

$$\sigma = [1 - (Fr/Fr_c)^m] \tag{76}$$

which multiplies the first two (inertia) terms in the momentum equations [(30) and (51)]. Fr is the Froude number of the flow in any ith Δx reach, and the exponent (m) varies from 1 to 10, with 5 usually preferred, and $0.85 < Fr_c < 0.95$ is the specified range for Fr_c. The filter takes on a value of zero when Fr $\geq$ 1. The local partial inertia filter (σ) avoids numerical difficulties associated with mixed flows while introducing negligible errors, less than about 1 to 2% for all flow conditions.

Flow Through a System of Rivers

A river system consisting of a main-stem river and one or more tributaries is efficiently solved using an iterative relaxation method (Fread, 1973, 1985) in

which the flow at the confluence of the main-stem and tributary is treated as the lateral inflow/outflow (q) in Eqs. (29) and (30). This algorithm was extended so as to treat a dendritic system of waterways having nth-order tributaries (Lewis et al., 1996) and further extended to treat a river system that has any bifurcations such as islands, along with or without nth-order tributaries (Jin, et al., 2000). Also, a less versatile direct solution technique can be used (Fread, 1985), wherein three internal boundary equations conserve mass and momentum at each bifurcation or junction confluence. The resulting system of algebraic equations uses a special sparse matrix Gaussian elimination technique for an efficient solution (Fread, 1983).

REFERENCES

Abbott, M. B., *An Introduction to the Method of Characteristics*, American Elsevier, New York, 1966.

Amein, M., and C. S. Fang, Implicit flood routing in natural channels, *J. Hydraul. Div., ASCE*, *96*, 2481–2500, 1970.

Arcement, G. J., Jr., and V. R. Schneider, *Guide for Selecting Manning's Roughness Coefficients for Natural Channels and Flood Plains*, Report No. RHWA-TS-84-204, U.S. Geological Survey for Federal Highway Administration, National Technical Information Service, PB84-242585, 1984.

Baltzer, R. A., and C. Lai, Computer simulation of unsteady flow in waterways, *J. Hydraul. Div. ASCE*, *94*, (HY4), 1083–1117, 1968.

Barkow, R. L., *UNET One-Dimensional Unsteady Flow Through a Full Network of Open Channels, Users Manual*, Hydrologic Engineering Center, U.S. Army Corps of Engineers, Davis, CA, 1990.

Barnes, Jr., H. H., *Roughness Characteristics of Natural Channels*, Geological Survey Water-Supply Paper 1849, U.S. Government Printing Office, Washington, DC, 1967.

Boussinesq, J., Theory of the liquid intumesence, called a solitary wave or a wave of translation, propagated in a channel of rectangular cross section, *Comp. Rend. Acad. Sci.*, *72*, 755–759, 1871.

Chow, V. T., *Open-Channel Hydraulics*, McGraw-Hill, New York, 1959.

Chow, V. T., *Handbook of Applied Hydrology*, Sections 7 and 25-II, McGraw-Hill, New York, 1964.

Chow, V. T., D. R. Maidment, and L. W. Mays, *Applied Hydrology*, McGraw-Hill, New York, 1988.

Cunge, J. A., On the subject of a flood propagation computation method (Muskingum method), *J. Hydraul. Res.*, *7*, (2), 205–230, 1969.

Cunge, J. A., F. M. Holly, Jr., and A. Verway, *Practical Aspects of Computational River Hydraulics*, Pitman, Boston, MA, 1980.

DeLong, L. L., Extension of the unsteady one-dimensional open-channel flow equations for flow simulation in meandering channels with flood plains, in *Selected Papers in Hydrologic Science*, U.S. Geological Survey Water Supply Paper 2220, 1986, pp. 101–105.

DeLong, L. L. Mass conservation: 1-D open channel flow equations, *J. Hydraul. Div.*, *115*(2), 263–268, 1989.

Dooge, J. C. I., W. G. Strupczewski, and J. J. Napiorkowski, Hydrodynamic derivation of storage parameters of the Muskingum model, *J. Hydrol.*, (54), 371–387, 1982.

Dronkers, J. J., Tidal computations for rivers, coastal areas, and seas, *J. Hydraul. Div., ASCE*, *95*(HY1), 29–77, 1969.

Fread, D. L., Discussion of implicit flood routing in natural channels, by M. Amein and C. S. Fang, *J. Hydraul. Div. ASCE*, *97*(HY7), 1156–1159, 1971.

Fread, D. L., Technique for implicit dynamic routing in rivers with tributaries, *Water Resour. Res.*, *9*(4), 918–926, 1973.

Fread, D. L., *Numerical Properties of Implicit Four-Point Finite Difference Equations of Unsteady Flow*, HRL-45, NOAA Technical Memo NWS HYDRO-18, Hydrologic Research Laboratory, National Weather Service, Silver Spring, MD, 1974.

Fread, D. L., The development and testing of a dam-break flood forecasting model, in *Proceedings of Dam-Break Flood Modeling Workshop*, U.S. Water Resources Council, Washington, DC, 1977, 164–197.

Fread, D. L., NWS operational dynamic wave model, in *Verification of Mathematical and Physical Models*, *Proceedings of 26th Annual Hydr. Div. Specialty Conf.*, American Society of Chemical Engineers, College Park, MD, 1978, pp. 455–464.

Fread, D. L., Computational extensions to implicit routing models, in *Proceedings of the Conference on Frontiers in Hydraulic Engineering*, MIT, Cambridge, MA, 1983, pp. 343–348.

Fread, D. L., Channel routing, in M. G. Anderson and T. P. Burt (Eds.), *Hydrological Forecasting*, Wiley, New York, 1985, pp. 437–503.

Fread, D. L., *The NWS DAMBRK Model: Theoretical Background/User Documentation*, HRL-256, Hydrologic Research Laboratory, National Weather Service, Silver Spring, MD, 1988.

Fread, D. L., Flood routing and the Manning n, in B. C. Yen, (Ed.), *Proceedings of the International Conference for Centennial of Manning's Formula and Kuichling's Rational Formula*, Charlottesville, VA, 1989, 699–708.

Fread, D. L., Flow routing, in D. Maidment (Ed.), *Handbook of Hydrology* McGraw-Hill, New York, 1992, pp. 10.1–10.36.

Fread, D. L., Selection of Δx and Δt computational steps for four-point implicit non-linear dynamic routing models, in *Proceedings, National Hydraulic Engineering Conference*, American Society of Chemical Engineers, San Francisco, CA, 1993.

Fread, D. L., Dam-breach floods, in V. J. Singh (Ed.), *Hydrology of Disasters*, Kluwer Academic, Boston, 1995, pp. 85–126.

Fread, D. L., and T. E. Harbaugh, Open channel profiles by Newton iteration technique, *J. Hydrol.*, *13*, 79–80, 1971.

Fread, D. L. and J. M. Lewis, NWS FLDWAV Model: Theoretical Description/User Documentation, HAL-406, Hydrologic Research Laboratory, National Weather Service, Silver Spring, MD, Nov., 1998.

Fread, D. L., M. Jin, and J. M. Lewis, An LPI numerical implicit solution for unsteady mixed-flow simulation, in *Proceedings, North American Water and Environment Congress '96*, *American Society of Chemical Engineers*, Anaheim, CA, 1996.

Froehlich, D. C., Embankment-dam breach parameters, in *Proceedings of the 1987 National Conference on Hydraulic Engineering*, American Society of Chemical Engineers, New York, 1987, pp. 570–575.

Garrison, J. M., J. P. Granju, and J. T. Price, Unsteady flow simulation in rivers and reservoirs, *J. Hydraul. Div. ASCE*, *95*(HY5), 1559–1576, 1969.

Gray, W. G., G. F. Pinder, and C. A. Brebbia, *Finite Elements in Water Resources*, Pentech Press, London, 1977.

Henderson, F. M., *Open Channel Flow*, Macmillan, New York, 1966, pp. 285–287.

Hydrologic Engineering Center, *HEC-1 Flood Hydrograph Package—Users Manual*, U.S. Army Corps of Engineers, Davis, CA, 1981.

Isaacson, E., J. J. Stoker, and A. Troesch, *Numerical Solution of Flood Prediction and River Regulation Problems*, Report II/III, No. IMM-NYU-205/235, New York University Institute of Mathematics and Science, New York, 1954, 1956.

Jarrett, R. D., Hydraulics of high-gradient streams, *J. Hydraul. Div. ASCE*, *110*(HY11), 1519–1539, 1984.

Jin, M., and D. L. Fread, One-dimensional routing of mud/debris flows using NWS FLDWAV model, in *Proceedings, First International Conference on Debris-Flow Hazards Mitigation: Mechanics, Prediction, and Assessment*, American Society of Chemical Engineers, New York, 1997.

Jin, M., D. Fread, and J. Sylvestre, Channel routing in river networks using NWS FLDWAV model, in 2000 Joint Conference on Water Resources Engineering and Water Resources Planning and Mangement, ASCE Proceedings CD ROM.

Jones, S. B., Choice of space and time steps in the Muskingum-Cunge flood routing method, *Proc. Inst. Civ. Eng.*, Part 2, No. 71, 759–772, 1981.

Laplace, P. S., Recherches sur quelques points due systeme du monde, [Researches on some points of world system], in *Memoirs*, Vol. 9, Acad. Sci., Paris, 1776.

Lewis, J. M., D. L. Fread, and M. Jin, An extended relation technique for modeling unsteady flows in channel networks using the NWS FLDWAV model, in *Proceedings, North American Water and Environment Congress '96*, American Society of Chemical Engineers, Anaheim, CA, 1996.

Liggett, J. A., Basic equations of unsteady flow, in K. Mahmood and V. Yevjevich, (Eds.), *Unsteady Flow in Open Channels*, Water Resources, Fort Collins, CO, 1975, pp. 29–62.

Liggett, J. A., and J. A. Cunge, Numerical methods of solution of the unsteady flow equations, in K. Mahmood and V. Yevjevich (Eds.), *Unsteady Flow in Open Channels*, Vol. I, Water Resources, Fort Collins, CO, 1975, pp. 89–182.

Linsley, R. K., M. A. Kohler, and J. L. H. Paulhus, *Hydrology for Engineers*, McGraw-Hill, New York, 1986, pp. 502–530.

Manning, R., On the flow of water in open channels and pipes, *Trans. Inst. Civil Eng. Ireland*, *20*, 161–195, 1889.

McCarthy, G. T., The unit hydrograph and flood routing, in *Conf. of the North Atlantic Div.*, U.S. Corps of Engineers, New London, CT, 1938.

Miller, W. A., and J. A. Cunge, Simplified equations of unsteady flow, in K. Mahmood and V. Yevjevih (Eds.), *Unsteady Flow in Open Channels*, Vol. I Water Resources, Fort Collins, CO, 1975, pp. 183–257.

Newton, Sir I., Propositions, Book 2, in *Principia*, Royal Society, London, 1687, pp. 44–46.

O'Brien, J. S., and P. Julien, Physical properties and mechanics of hyper-concentrated sediment flows, in D. S. Bowles (Ed.), *Delineation of Landslide, Flash Flood, and Debris Flow*

Hazards in Utah, General Series UWRL/G-85/03, Utah State University, Utah Water Research Laboratory, Logan, UT, 1984, pp. 260–279.

Poisson, S. D., Memoir on the theory of waves, in *Memoirs*, Vol. 1, Acad. Sci., Paris, 1816, pp. 71–186.

Ponce, V. M., and V. Yevjevich, V. Muskingum-Cunge method with variable parameters, *J. Hydraul. Div. ASCE*, *104*(HY12), 1663–1667, 1978.

Preissmann, A., Propagation of translatory waves in channels and rivers, in *Proc. First Congress of French Assoc. for Computation*, Grenoble, France, 1961, pp. 433–442.

Rajar, R., Mathematical simultion of dam-break flow, *J. Hydraul. Div. ASCE*, *104*(HY7), 1011–1026, 1978.

Saint-Venant, Barré de, Theory of unsteady water flow, with application to river floods and to propagation of tides in river channels, in *Computes rendus*, Vol. 73, Acad. Sci., Paris, France, 1871, 148–154, 237–240. (Translated into English by U.S. Corps of Engineers, No. 49-g, Waterways Experiment Station, Vicksburg, MS, 1949.)

Samuels, P. G., *Models of Open Channel Flow Using Preissmann's Scheme*, Cambridge University Press, Cambridge, 1985, pp. 91–102.

Sayed, I., and D. C. Howard, Application of dynamic backwater modeling to Mactaquac headpond—Saint John River, N.B., in *Proceedings of 6th Canadian Hydrotechnical Conference*, Canadian Society for Civil Engineering, 1983, pp. 203–220.

Schaffranek, R. W., *Flow Model for Open Channel Reach or Network*, Professional Paper No. 1384, U.S. Geological Survey, 1987.

Singh, V. P., and R. C. McCann, Some notes on Muskingum method of flood routing, *J. Hydrol.*, *48*(3), 343–361, 1980.

Stoker, J. J., *Numerical Solution of Flood Prediction and River Regulation Problems; Derivation of Basic Theory and Formulation of Numerical Methods of Attack*, Report I, No. IMM-NYU-200, New York University Institute of Mathematical Science, New York, 1953.

Stoker, J., *Water Waves*, Interscience, New York, 1957, pp. 452–455.

Streeter, V. L., and E. B. Wylie, *Hydraulic Transients*, McGraw-Hill, New York, 1967, pp. 239–259.

Strelkoff, T., The one-dimensional equations of open-channel flow, *J. Hydraul. Div., ASCE*, *95*(HY3), 861–874, 1969.

Strelkoff, T., Numerical solution of Saint-Venant equations, *J. Hydraul. Div. ASCE*, *96*(HY1), 223–252, 1970.

Strelkoff, T., and N. D. Katopodes, Border irrigation hydraulics with zero inertia, *J. Irrig./Drain. Div. ASCE*, *103*, 325–342, 1977.

Strupczewski, W., and Z. Kundzewicz, Translatory characteristics of the Muskingum method of flood routing—a comment, *J. Hydrol.*, *98*, 363–368, 1980.

Viessman, Jr., W., J. W. Knapp, G. L. Lewis, and T. E. Harbaugh, *Introduction to Hydrology*, 2nd ed., Intext Educational Publishers, New York, 1977.

CHAPTER 31

HYDROLOGIC MODELING FOR RUNOFF FORECASTING

HOSHIN GUPTA

1 INTRODUCTION

The problem of forecasting streamflow levels given precipitation data has received the time and attention of a great many hydrologists. Models developed for this purpose have ranged from simple to extremely complex. The simplest ones are based on input–output regression-type relationships, while the most complex ones attempt to represent the detailed water and energy balance physics occurring in the watershed. The complex models are motivated largely by experimental evidence that the subwatershed-scale components of the rainfall–runoff process are strongly nonlinear, time variable, and spatially distributed. However, the processes of aggregation, attenuation, loss, and delay tend to result in an overall watershed response that is far less complex than the point-scale behavior. The effects of subwatershed-scale variability tend to be smoothed and poorly observable (to varying degrees) in the overall watershed-scale response. Thus, while remarkable progress has been made in understanding the physics of how precipitated water moves once it reaches the ground, the level of model complexity required to provide accurate runoff forecasts for any chosen watershed remains unclear. Even less clear is how this complexity varies with climatology, watershed size, and geologic and physiographic characteristics of the landscape.

2 MODELING AND COMPLEXITY

In the absence of such clarity, a wide variety of hydrologic models have found their way into the literature (Singh, 1995). The essential difference in these models is the

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

manner in which the underlying processes that transform precipitation into streamflow are conceptualized. The more complex models have been motivated by the scientific pursuit of knowledge and are based on painstaking research into the physics of subwatershed-scale hydrologic processes. Such models attempt, in particular, to account for the spatially and temporally varying nature of watershed inputs (precipitation, solar radiation, etc.), losses (evapotranspiration), and characteristics (topography, permeability, vegetation, etc.). We shall refer to this modeling approach as "physics based." Perhaps the most well-known exponent of this approach is the Systéme Hydrologique Européen (SHE) model (Abbott, 1986). More recent developments are the soil–vegetation–atmosphere–transfer schemes (SVATS) used for climate studies, such as Biosphere Atmosphere Transfer Scheme (BATS) (Dickinson, 1993), Simple Biosphere Model 2 (SiB2) (Randall, 1996), and Variable Infiltration Capacity 2-Layer Model (VIC-2L) (Liang, 1994).

At the other end of the spectrum, the simplest models have been motivated by engineering considerations based on a real need to provide quick and accurate forecasts of streamflow levels in the simplest possible way, particularly wherever human interests are at stake (such as flood-prone locations). Such models attempt to establish direct regression-like relationships between the input and output time series; generally, the streamflow value is regressed on values of precipitation and streamflow at previous times. We shall refer to this modeling approach as "systems theoretic." The most popular systems-theoretic methods have been the ARMAX (auto-regressive moving average with exogenous inputs) (Box, 1976; Salas, 1980) and the ANN (artificial neural network) (Hsu, 1995, 1997).

A third category of models, which are of intermediate complexity, are based on attempts to conceptualize the simplified (lumped) watershed-scale behavior resulting from the integrated effect of the subwatershed-scale hydrologic processes. Such models typically use simple linear and nonlinear tank components (reservoirs) to represent the primary soil moisture zones in the watershed and describe the manner in which moisture exchanges among these stores take place. We shall refer to such models as "conceptual." It is important to note that such models are based on *speculative conjecture* as to how best to partition the watershed into components and how to represent the integrated behavior of each component. This, and the fact that conceptual models are relatively simple to program into a computer, has encouraged a great deal of intellectual experimentation, resulting in a proliferation of conceptual models with widely differing structures. At the simple end, we have methods such as the API (antecedent precipitation index) and UHG (unit hydrograph) which, in a simple manner, partition the watershed response into precipitation excess and infiltration (based on an antecedent soil moisture index) and use linear equations to transform the precipitation excess into streamflow forecasts. Models at the intermediate level include the HEC-1 model (U.S. Army Corps of Engineers, 1973, 1985). At the complex level, we have methods such as the Stanford watershed model (SWM) (Crawford, 1966), the Institute of Hydrology Model (IHDM) (Beven, 1987), the Kineros model (Woolhiser, 1990), the Sacramento soil moisture accounting model (SAC-SMA) (Burnash, 1973), and TOPMODEL (Beven, 1979) that have numerous components. Within the United States, the most widely used of

these may well be the API, UHG, and SAC-SMA models because they are extensively used by various regional offices of the U.S. National Weather Service for flood forecasting. Such models are currently being built into more general "modeling systems" such as the advanced hydrologic prediction system (AHPS) of the U.S. National Weather System and the modular modeling system (MMS) of the U.S. Geological Survey. These systems allow the user to build up a complete model by selecting the components from libraries containing several alternative conceptual representations.

Finally, the last half-decade has seen the emergence of a subclass of conceptual models that seek to strike a reasonable and parsimonious balance between the three issues of (a) scientific understanding (physics), (b) speculative conjecture about the nature of integrated watershed-scale processes (conceptualization), and (c) the level of model complexity that can actually be supported by the available watershed response data (i.e., the systems-theoretic issues of observability and identifiability). Examples of such models are the IHACRES [see e.g., Jakeman (1990, 1993)] model and the related HyMod under development at the University of Arizona (Boyle, 2000). For want of a better terminology, we follow Wheater (1993) in referring to these as "hybrid" models.

The three mechanisms of scientific understanding, conceptualization, and data-supportable complexity can be likened to the legs of a stool that must be of proper and complementary length so that the sitting surface is balanced and can perform its intended function (Fig. 1). The key issue in selecting an appropriate model is this intended function. A physics-based model such as SHE and some conceptual models such as TOPMODEL and Kineros may be clearly appropriate for detailed watershed modeling and for testing hypotheses about watershed behavior under perturbed conditions. On the other hand, Hsu et al. (1995) have shown that simple ANN-type systems-theoretic models can give one-step-ahead forecasts that are more accurate than those given by conceptual models, while requiring relatively minor computational resources and being quick and easy to build. However, if the intended

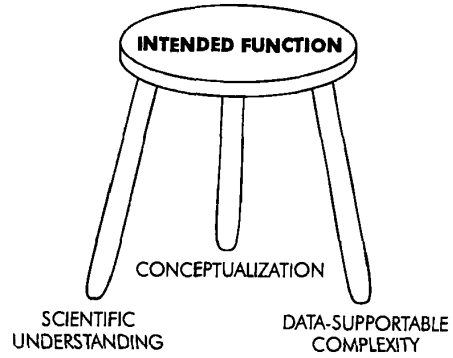


Figure 1 Issues influencing model development and selection.

function is *both* accurate operational streamflow forecasting *as well* as insight into evolving watershed behavior, the emerging evidence suggests that hybrid models such as IHACRES and HyMod, which merge the strengths of the conceptual and systems-theoretic approaches, may prove to be the optimal choice.

3 MODEL PARAMETER ESTIMATION, CALIBRATION, AND EVALUATION

The model selected must be made specific to a watershed by estimating values for its parameters. In the case of physically based models and some conceptual models, approximate values (or ranges of values) for many of the parameters can sometimes be estimated from maps or field measurements. However, because all such models involve conceptualization (simplification and distortion from reality), the parameter estimates obtained in this manner can invariably be improved by calibration to historical input–output data. Certainly, in the case of systems-theoretic models, the only method for inferring structural complexity and parameter values is an automated computer-based identification procedure. Because each of the available systems-theoretic modeling approaches (such as ARMAX and ANN) is generally accompanied by clear procedures for model building and parameter estimation, they will not be described here. The discussion here will focus on parameter estimation for the other three categories of models via a procedure called model calibration.

The model calibration process involves five interrelated components: (a) data set, (b) constraints, (c) measures of closeness, (d) parameter adjustment procedure, and (e) evaluation procedure. Each of these components is discussed in turn.

Data Set

The input–output data set to be used for inferring model parameters must be carefully selected from the historical record to be representative of the behavior of the watershed. Two issues are important here—data quality and data quantity. Data quality has two subissues that must be considered. The first is simply that the data must be checked for accuracy and reliability (i.e., errors in measurement and/or recording). To take a trivial example, if the precipitation records indicate a large storm event but the flow records do not show a response (or vice versa), we might suspect the accuracy of the data. The second subissue is related to data informativeness; i.e., the data must be representative of the important characteristic modes of watershed behavior. For example, if the purpose of the model is flood forecasting, the data must certainly contain several significant storm events. These data will provide information about the parameters related to the partitioning of precipitation into flow components having different recession rates. However, the data must also contain several representative interstorm periods so that information regarding the parameters controlling streamflow recession as well as rates of evaporation loss can be deduced. There have been only a few studies investigating this issue. Gupta (1985) used a theoretical analysis to show that "threshold-type"

parameters are best identified when the data are selected to ensure that the model behavior tends to switch across the threshold numerous times; surprisingly, the amount of time spent in each mode of behavior is largely irrelevant. Yapo (1996) studied the reliability of parameter estimates of a conceptual rainfall–runoff model using 40 years of data and clearly demonstrated that the most reliable results are provided by using "wet" years for calibration.

The aforementioned studies also addressed the issue of data quantity (length). Gupta (1985) showed theoretically that a (daily) data set of approximately 3 years length is desirable for model calibration, and that additional amounts of data will provide only marginal gains *unless containing significantly new information*. Yapo (1996) found, however, that for the Leaf River in Mississippi, the SAC-SMA conceptual model requires at least 8 to 10 years of data for reliable calibration results to be obtained, suggesting that the variability of information in a hydrologic data set may extend over approximately a decade.

Having selected the calibration period data set, the next important decision is the selection of an appropriate length for the "buffer" period. A buffer period is a short data segment at the very beginning of the data set for which the measures of closeness (see below) are not computed. The intention is to minimize any potential bias in the calibration procedure caused by uncertain initialization of the model state variables. Because a watershed model tends to average and attenuate inputs, it will also attenuate the impact of initialization errors over time. A buffer period of 90 to 180 days beginning near the end of a long recession and approximately a week or two before the end of the dry season seems to be a good choice.

Constraints

The search for a better (or "best") parameter set is facilitated greatly by specifying upper and lower limits for each of the parameters—this defines the "feasible" region in which an "optimal" parameter set is expected to lie. A useful perspective is to consider this feasible region to be the initial uncertainty in the parameter estimates, based on available prior information. For certain parameters, these upper and lower limits are easily selected based on physical considerations related to the characteristics of the watershed. For parameters such as equation exponents defining the degree of nonlinearity of a transformation, one may only be able to guess at an appropriate range of values. The selection of constraints is rather model dependent and can be quite subjective. However, if the parameter adjustment procedure is sufficiently powerful, the impact of this subjectivity should be minimal.

Measures of Closeness

To select a better (or best) parameter set from the feasible parameter space, we must be able to compare and evaluate, in some manner, the model performance associated with different parameter sets. The indicator of model performance is usually taken to be a comparison between the observed (measured) watershed output time series and the corresponding model simulated quantities. In general, particularly for runoff

forecasting models, the indicator representing watershed behavior is the sequence of observed streamflow levels at the watershed outlet, although streamflow level data at gauging stations within the watershed or soil moisture data at specific points within the watershed may sometimes also be available. Other models may have multiple outputs that can serve as indicators of model behavior. In the case of watershed hydrochemical models, we may have data on concentrations of various chemical species, and in the case of watershed scale water-and-energy budget models [e.g., Dickinson (1993) and Schaake (1996)], we may have data on surface soil temperature, emitted short- and long-wave radiation, sensible and latent heat fluxes, etc.

The million-dollar question is: What method should be used to compare the model-simulated streamflow values and the observed streamflow data? A visual comparison of the two time series plotted together on the same graph is intuitively appealing (Fig. 2) but is made difficult by the large number (say n) of time steps at which the simulated streamflow values must be compared to the observed data. If $E_t = \{e_t = s_t - o_t,\ t = 1, \ldots, n\}$ represents the vector of differences between each simulated flow (s_t) and its corresponding observed data value (o_t), the method of visual comparison will involve adjusting the parameters to simultaneously make each one of the e_t differences as small as possible. Because this approach is subjective, different hydrologists will tend to judge different model-simulated time series (and their associated parameter sets) as being better. Further, while it may be relatively simple to decide that a certain approximate region of the parameter space gives better simulations than some other regions of the parameter space, it can be very difficult to narrow down the choice (see section below on parameter adjustment). The method of visual comparison is also difficult if not impossible to automate.

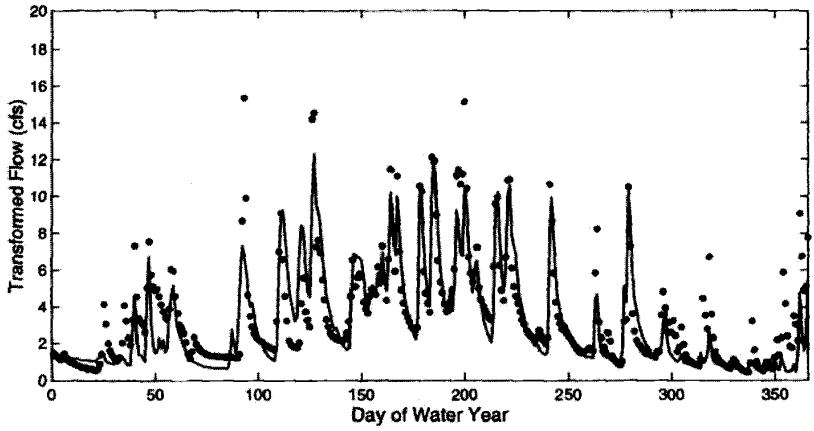


Figure 2 Plot of observed flow (·) vs. model simulated flow (−). Values are in the transformed space (to observe behavior in the full range of flows better), where transformed flow = $[(\text{flow} + 1)\lambda - 1]/\lambda$, and $\lambda = 0.3$.

An alternative to visual comparison is to define a mathematical measure of the "size" of the vector E_t. However, there are an infinite number of ways in which this can be done. The most popular implementation is to compute a scalar measure of the average size of the differences, such as the mean-squared error [MSE = mean(e_t^2, $t = 1, \ldots, n$)) or the mean absolute error (MAE = mean($|e_t|$, $t = 1, \ldots, n$)]. A number of different such measures, which are commonly called "objective" functions, have been suggested in the literature; Table 1 lists many of the measures used by the U.S. National Weather Service for calibration of its flood forecast models. A related approach is to treat the residuals as though they have stochastic properties and belong to some preassumed probability distribution, usually assumed to be Gaussian. Under this assumption, it is possible to develop maximum-likelihood (ML) measures having theoretical underpinnings; for example, the heteroscedastic maximum-likelihood estimator [HMLE = mean ($w_t^* e_t^2$, $t = 1, \ldots, n$); $w_t = o_{\text{obs},t}^{2(\lambda-1)} / (\prod_1^n o_{\text{obs},t}^{2(\lambda-1)})^{1/n}$; λ is a parameter to be estimated] criterion developed by Sorooshian (1980) assumes that the residuals are Gaussian, uncorrelated, unbiased, and have

TABLE 1 Objective Functions Used by National Weather Service for Calibration of SAC-SMA Model

Name	Description	Formula	
DRMS	Daily root-mean-squared error	Minimize w.r.t θ	$\sqrt{\frac{1}{n}\sum_{t=1}^{n}[s_t - o_t(\theta)]^2}$
TMVOL	Total mean monthly volume-squared error	Minimize w.r.t θ	$\sum_{i=1}^{n\text{month}} \left\{ \frac{1}{n\text{day}(i)} \sum_{-1}^{n\text{day}(i)} [s_t - o_t(\theta)] \right\}^2$
ABSERR	Mean absolute error	Minimize w.r.t θ	$\frac{1}{n}\sum_{t-1}^{n} \lvert s_t - o_t(\theta) \rvert$
ABSMAX	Maximum absolute error	Minimize w.r.t θ	$\max_{1 \le t \le n} \lvert s_t - o_t(\theta) \rvert$
NS	Nash–Sutcliffe measure	Minimize w.r.t θ	$1 - \frac{\frac{1}{n}\sum_{t=1}^{n}[s_t - o_t(\theta)]^2}{\frac{1}{n}\sum_{t=1}^{n}(s_t - \bar{s})^2}$
BIAS	Bias (mean daily error)	Minimize w.r.t θ	$\frac{1}{n}\sum_{t=1}^{n}[s_t - o_t(\theta)]$
PDIFF	Peak difference	Minimize w.r.t θ	$\max_{1 \le t \le n}\{s_t\} - \max_{1 \le t \le n}\{o_t(\theta)\}$
RCOEF	First lag auto-correlation	Minimize w.r.t θ	$\frac{\frac{1}{n}\sum_{t=1}^{n}[s_t - o_t(\theta)][s_{t+1} - o_{t+1}(\theta)]}{\sigma_s \sigma_{0(\theta)}}$
NSC	Number of sign changes	Minimize w.r.t θ	(Count the number of times the sequence of residuals changes sign)

nonhomogenous variance. The advantage of the ML approach is that the validity of the underlying assumptions can be verified by a postcalibration residual analysis. It is important to note that each of these scalar measures defines a different way to gauge the "size" of the error vector, and the minimum value for each will define a simulated streamflow sequence that is "close" to the observed data in a different way. If a certain scalar measure is selected, then it is possible (in principle) to find a single parameter set (or a small region of the feasible space) that minimizes that measure. This makes the model calibration procedure much easier to automate so that the accuracy, speed, and efficiency of a computer can be exploited. Nonetheless, while more efficient than visual comparison, the use of scalar measures is perhaps no less subjective.

Parameter Adjustment Procedure

The parameter adjustment procedure is a directed trial-and-error process by which the parameters are iteratively adjusted to move the model behavior closer to the observed data. The choice of procedure is related to the measure of closeness selected (see above). If the calibration is performed by an expert hydrologist having a great deal of familiarity with the nuances of the model, the method of manual parameter adjustment guided by visual comparison can be extremely effective. However, manual calibration has several drawbacks. First, the procedure requires a great deal of subtlety in evaluating the visual goodness of fit, something that takes time and training to develop. Even to the trained eye, there may appear to exist numerous equally "good" parameter sets that are difficult to distinguish (Beven, 1992; Freer, 1996). Different good parameter sets will appear to match the data well in different ways, and moving from one set to another will trade-off an improvement in matching some parts of the data against deterioration in matching other parts of the data (Gupta, 1998). In practice, the calibration expert can also support the qualitative visual comparison with one or more quantitative scalar measures (e.g., see Table 1). However, this evaluation process still tends to be greatly complicated by the large number of model parameters to be adjusted and their tendency to have interacting and compensating effects on the output. Furthermore, the process can be very time intensive, particularly when the model contains numerous subcomponents and a large number of parameters (e.g., manual calibration of the SAC-SMA model can take several person-days of dedicated effort). These difficulties tend to limit widespread utility of the more complex and sophisticated models.

An alternative to manual parameter adjustment is to use the speed and power of a computer to automatically search the feasible parameter space for "better" solutions. In this approach, the measure of closeness is typically one of the scalar measures of closeness described earlier. A great deal of research has gone into the development of an automatic parameter adjustment procedure that gives satisfactory results while being reliable (effective) and efficient. A satisfactory result is one that gives model simulations similar to those obtained by an expert manual calibration, while resulting in parameter estimates that are conceptually realistic; a reliable procedure is one

TABLE 2 Summary of Five Major Characteristics Complicating the Optimization Problem in CRR Model Calibration

1. Regions of attraction	More than one main convergence region
2. Minor local optima	Many small "pits" in each region
3. Roughness	Rough response surface with discontinuous derivatives
4. Sensitivity	Poor and varying sensitivity of response surface in region of optimum, and nonlinear parameter interaction
5. Shape	Nonconvex response surface with long curved ridges

that consistently provides satisfactory results; and an efficient procedure is one that requires only small amounts of computer time.

The earliest attempts at automatic calibration drew on a class of function optimization techniques called "local search" procedures; examples include the pattern search method (Hooke, 1961), the rotating directions method (Rosenbrock, 1960), the downhill simplex method (Nelder, 1965), and various versions of the Gauss–Newton quadratic approximation method (Luenberger, 1984). It quickly became apparent that such methods were highly unreliable; independent trials of the algorithm initiated from different initial parameter estimates would converge to widely differing solutions. A study by Duan (1993) demonstrated conclusively the reasons for this poor performance; the response surface of the scalar measure being optimized typically has several characteristic properties (see Table 2) that local search

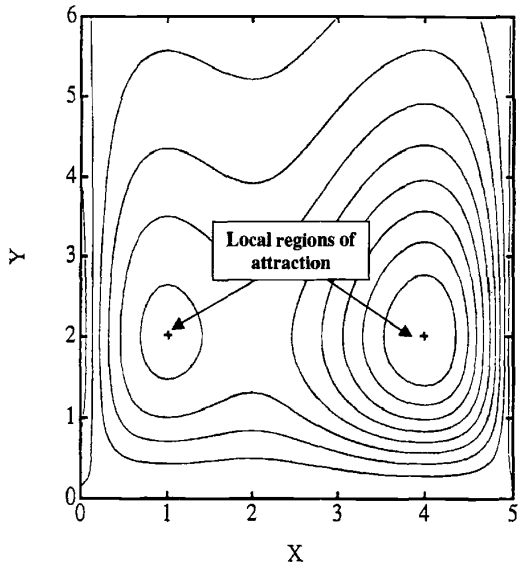


Figure 3 Function response surface showing multiple regions of attraction.

algorithms are not able to handle well. The most important of these are the existence of more than one primary region of attraction (see Fig. 3) as well as large numbers of local optima throughout the feasible space (see Fig. 4). The focus therefore shifted to trying the existing "global search" methods including adaptive random search (Brazil, 1987), the genetic algorithm (Wang, 1991; Tanakamaru, 1995), and the multistart simplex (Duan, 1992; Gan, 1996). The most successful method to date has been the shuffled complex evolution (SCE-UA) method recently developed at the University of Arizona (Duan, 1992, 1994; Sorooshian, 1993), which has proved to be both reliable and relatively efficient (see Fig. 5).

It should be noted that "manual" and "automatic" parameter adjustment approaches have mutually complementary strengths and weaknesses, which suggests the implementation of a hybrid approach that draws on the strengths of each (while minimizing their weaknesses). The strength of the manual approach is its ability, when successful, to provide very satisfying model calibrations because visual comparisons draw on the human ability to perceive patterns that are not easy to detect using numerical techniques. The strength of the automatic approach is that it can very quickly and rapidly find the region(s) of the parameter space that give relatively close matching of the simulated flows and observed data, while manipulating large (even bewildering) numbers of mutually compensating and interacting parameters. The hybrid procedure therefore involves two steps. In the first step, the automatic procedure is used to quickly find several solutions that seem to have similar ability to match the data when measured using one *or more* of the scalar

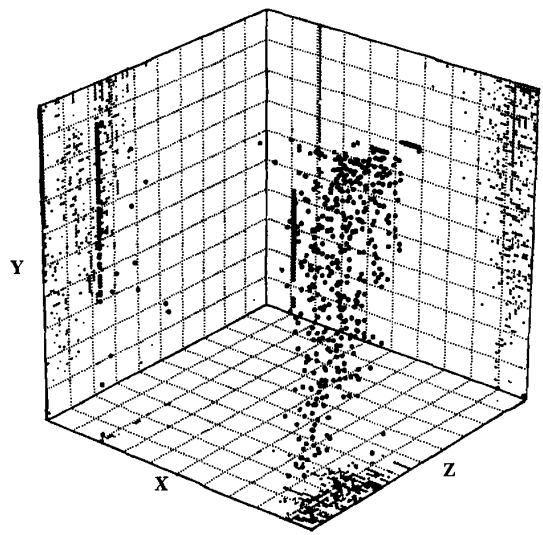


Figure 4 Locations of local optima in three-dimensional parameter subspace; each dot represents optima for a local region.

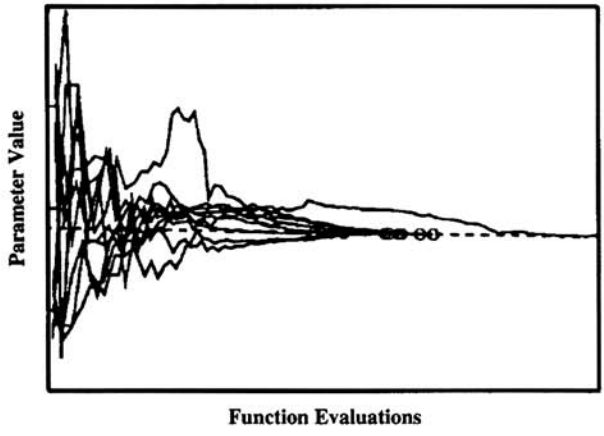


Figure 5 Convergence of model parameter for 10 different trials of the SCE-UA method; dotted line represents the true paramenter value.

numerical measures of closeness described earlier. These then become the starting point for a manual procedure of refinement in which the expertise of the hydrologist can be used to further improve the solution. The computer does what it does better than a human, which is to search through large numbers of options very quickly and reject the unacceptable ones. The human does what humans do better than a computer, which is to use perceptual discrimination to make qualitative distinctions that are difficult to describe mathematically.

Developments of this hybrid approach through the use of multi-objective procedures can be found in the work of Gupta (1998) and Yapo (1998). The multiobjective global optimization strategy, MOCOM (multi-objective complex evolution), is used to identify the set of solutions that provide a "trade-off" in simultaneously minimizing several criteria that measure the goodness of fit of the model to the calibra-

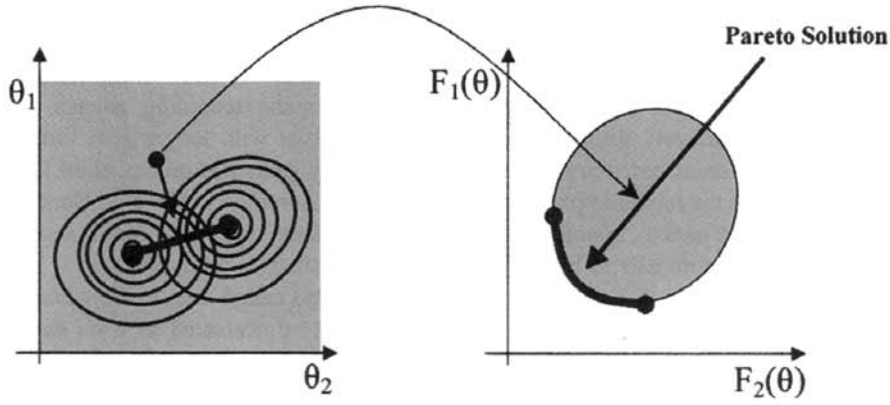


Figure 6 Identification of trade-off solutions using multiobjective optimization strategy.

tion data (Fig. 6). The hydrologist can then use visual means to identify the most perceptually appealing solution(s).

Evaluation Procedure

Once a model has been calibrated by one of the methods outlined above, it is useful to evaluate the result by testing its performance using data not employed for model calibration. For flood forecast models, one might (if possible) select a period of data of comparable length to the calibration period containing several significant storm events. Visual comparison of the observed and simulated outputs for this evaluation period and a check of the goodness-of-fit statistics can reveal any obvious divergence of the model performance from reality. This can also give a reasonable estimate of the approximate forecasting accuracy that can be expected when the model is used for real-time forecasting. Particular attention should be given to any tendencies for the model simulations to be biased at different streamflow levels. Admittedly, the process of model evaluation is somewhat subjective; however, if the simulation performance over the evaluation period is essentially similar to that over the calibration period, the model can then be used for flood forecasting with some understanding of its expected level of performance.

4 FORECASTING AND STATE UPDATING

The calibrated model can be implemented for real-time flood forecasting. The main issue here is that of the "lead time" (duration between time of making the forecast and time of actual occurrence). Clearly, the benefit of a flood forecast lies both in its accuracy *and* in its being available as early as possible before the actual event occurs. With this in mind, the model time step Δt must necessarily be shorter than the time of concentration of the watershed so that the precipitation data available up to time t are used to compute the model-simulated streamflow at time $t + \Delta t$; this is called a one-step-ahead forecast. For small watersheds, this time step Δt may be on the order of only a few hours, while for larger watersheds the time step may be one day or more. To maximize the forecast lead time, it is desirable that the precipitation measurement be either phoned or radioed in to the forecast center within minutes of its occurrence, or even telemetered in by automatic recording gauges and processed immediately through the model. If a forecast with longer lead time is required, it becomes necessary to obtain independently generated precipitation forecasts to feed to the model in place of precipitation measurements. The U.S. National Weather Service uses a "quantitative precipitation forecast" system to enable several time-step-ahead forecasts to be made for many watersheds (Funk, 1991).

A second issue is that of model state updating (also called data assimilation or filtering). Because the accuracy of each forecast can be evaluated as soon as the observed flow for that time step becomes available, this information should, in principle, be useful for adjusting the internal model states to maximize the accuracy of the next forecast. For example, underprediction of the observed flow may indicate that the model storages that represent the wetness of the various watershed compo-

nents are too dry and should be adjusted accordingly. Kitanidis (1980a, 1980b) rewrote the SAC-SMA model in a state-space form and implemented an extended Kalman filter to enable the model to correct for data errors. Because of the mathematical complexity of reworking a model into a state-space form, state updating has not become widely popular for use with flood forecast models. As the use of the simpler "hybrid" models becomes popular, we can expect to see more exploitation of systems-theoretic methods such as state updating to improve the performance of watershed models.

Finally, it is important to consider the forecasting uncertainty associated with the uncertainty in the model structure and parameter estimates. Some interesting (and somewhat similar) Monte Carlo approaches for representing forecast uncertainty include the generalized likelihood uncertainty estimation (GLUE) method [see, e.g., Beven (1992) and Freer (1996)], the Monte Carlo set membership (MCSM) procedure [see, e.g., Keesman (1990) and van Straten (1991)], and the prediction uncertainty (PU) method [see, e.g., Klepper (1991)]. For example, the GLUE procedure estimates the range of forecast uncertainty by estimating the likelihood associated with the individual forecasts given by different "equifinal" parameter sets in the feasible space.

5 EMERGING DIRECTIONS

It is the thesis of this chapter that the trend in hydrologic modeling for runoff forecasting will be toward a successful marriage of hydrologic science and systems theory, implemented through the coupling of hybrid watershed models with automated procedures for calibration and data assimilation for state updating. We can expect to see clear and rapid progress in all three of these components. Experiments with data from numerous watersheds will help in establishing general guidelines about the level of conceptual detail required to model the dominant watershed responses that are observable in the input–output data. The development of multiobjective calibration procedures (Gupta, 1998; Yapo, 1998) has already begun to merge the strengths of the manual and automated calibration procedures into an effective hybrid calibration method. The simplicity of the hybrid model structures will enable approximate Kalman filtering methods (or other uncertainty estimation methods) to be implemented for improving online forecasts. In addition, radar-based precipitation estimates are already replacing gage-based data and will encourage the development of "distributed" structures but parsimoniously parameterized hybrid watershed models. Finally, because the hybrid modeling approach provides us with a simple functional representation of the watershed, we can also expect progress in understanding how to apply watershed models to ungaged basins.

REFERENCES

Abbott, M. B., J. C. Bathurts, J. A. Cunge, P. E. O'Connell, and J. Rasmussen, An introduction to the European Hydrological System-Systeme Hydrologique Eurorpeen, ASHE@: 2.

Structure of a physically-based, distributed modeling system. *Journal of Hydrology, 87*, 61–77, 1986.

Beven, K. J., and M. Kirby, A physically based variable contributing area model of basin Hydrology. *Hydrological Sciences Bulletin, 24*, 43–69, 1979.

Beven, K., A. Calver, and E. Morris, The institute of hydrology distributed model. U.K. Institute of Hydrology Report No. 98, 1987.

Beven, K.J., and A.M. Binley, The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes, 6*, 279–298, 1992.

Brazil, L. E., and W. F. Krajweski, Optimization of complex hydrologic models using random search methods. Paper presented at Conference on Engineering Hydrology, Hydraulics Division, American Society of Civil Engineering, Williamsburg, Virginia, Aug. 3–7, 1987.

Box, G. E. P., and G. M. Jenkins, Time Series Analysis: Forecasting and Control. Holden-Day Inc., San Francisco, CA, 1976.

Boyle, D. P., H. V. Gupta, and S. Sorooshian, Toward improved calibration of hydrological models: combining the strengths of manual and automatic methods, *Water Resources Research, 36*, 12, 3663–3674.

Burnash, R. J. C., R. L. Ferrell, and R. A. McGuire, *A Generalized Streamflow Simulation System*, Jr. Fed-State River Forecast Center, Sacramento, CA, 1973, 204 pp. 1973.

Crawford, N. H., and R. K. Linsley, Digital Simulation in Hydrology: Stanford Watershed Model IV, Technical Report No. 39, Stanford University Dept. of Civil Engineering, 1966.

Dickinson, R. E., A. Henderson-Sellers, and P. J. Kennedy, Biosphere Atmosphere Transfer Scheme (BATS) Version le as coupled to the NCAR Community Climate Model. NCAR Technical Note, NCAR/TN-387+STR, National Center for Atmospheric Research, Boulder, CO, 1973, 72pp.

Duan, Q., V. K. Gupta, and S. Sorooshian, Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research, 28*, 4, 1015–1031, 1992.

Duan, Q., V. K. Gupta, and S. Sorooshian, A shuffled complex evolution approach for effective and efficient global minimization, *Journal of Optimization Theory and Applications, 76*, 3, 501–521, 1993.

Duan, Q., S. Sorooshian, and V. K. Gupta, Optimal use of the SCE-UA global optimization method for calibrating watershed models, *Journal of Hydrology, 158*, 265–284, 1994.

Freer, J., A. M. Beven, and B. Ambroise, Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resources Research, 32*, 7, 2161–2173, 1996.

Funk, T. W., Forecasting techniques utilized by the Forecast Branch of the National-Meteorological-Center during a major convective rainfall event, *Weather and Forecasting, 6*, 4, 548–564, Dec 1991.

Gan, T. Y. and G. F. Biftu, Automatic calibration of conceptual rainfall-runoff models: Optimization algorithms, catchment conditions, and model structure, *Water Resources Research*, 1996

Gupta, V. K., and S. Sorooshian, The relationship between data and the precision of parameter estimates of hydrologic models, *Journal of Hydrology, 81*, 57–77, 1985.

Gupta, V. K., S. Sorooshian, and P. O. Yapo, Towards improved calibration of hydrologic models: Multiple and non-commensurable measure of information, *Water Resources Research, 34*, 4, 751–763, 1998.

Hooke, R., and T. A. Jeeves, Direct search solutions of numerical and statistical problems. *Journal Assoc. Computer Mach., 8*, 2, 212–229, 1991.

Hsu, K., H. V. Gupta, and S. Sorooshian, Artificial Neural Network modeling of the rainfall-runoff process, *Water Resources Research*, *31*, 10, 2517–2530, 1995.

Hsu, K., C. Gao, S. Sorooshian, and H. V. Gupta, Precipitation estimation from remotely sensed information using artificial neural networks, *Journal of Applied Meteorology*, *36*, 9, 1176–1190, September, 1997.

Jakeman, A. J., I. G. Littlewood, and P. G. Whitehead, Computation of the instantaneous unit-hydrograph and identifiable component flows with application to 2 small upland catchments, *Journal of Hydrology*, *117*, 1-4, 275–300, September, 1990.

Jakeman, A., and G. Hornberger, How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, *29*, 8, 2637–2649, 1993.

Keesman, K. J., Set theoretic parameter estimation using random scanning and principal component analysis, *Mathematical Computation and Simulation*, 1990.

Kitanidis, P. K., and R. L. Bras, Adaptive filtering through detection of isolated transient errors in rainfall-runoff models, *Water Resources Research*, *16*, 4, 740–748, 1980a.

Kitanidis, P. K., and R. L. Bras, Real-time forecasting with a conceptual hydrological model: 1. Analysis of uncertainty, *Water Resources Research*, *16*, 6, 1025–1033, 1980b.

Klepper, O., H. Scholten, and J. P. G. van de Kamer, Prediction uncertainty in an ecological model of the Oosterschelde Estuary, *Journal of Forecasting*, *10*, 191–209, 1991.

Luenberger, D. G., Introduction to linear and nonlinear programming. Addison-Wesley, Menlo Park, CA, 1984.

Liang, X., and P. D. Lettenmaier, A simple hydrologically based model of land surface water and energy fluxes for general circulation models, *J. of Geophys. Res*, *99*, D7, 14, 415–14, 428, July, 1994.

Nelder, J. A., and R. Mead, A simplex method for function minimization, *Computer Journal*, *7*, 4, 308–313, 1965.

Randall, D. A., S. A. Dazlich, C. Zhang, A. S. Denning, P. J. Sellers, C. J. Tucker, L. Bounoua, S. O. Los, C. O. Justice, and I. Fung, A revised land surface parameterization (SiB2) for GCMs: 3. The greening of the Colorado State University general circulation model, *Journal of Climate*, *9*, 4, 738–763, April, 1996.

Rosenbrock, H. H., An automatic method of finding the greatest or least value of a function, *Computer Journal*, *3*, 175–184, 1960.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, WRP, Littleton, CO, 1960.

Schaake, J. C., V. I. Koren, and Q. Y. Duan, Simple water balance model for estimating runoff at different spatial and temporal scales, *Journal of Geophysical Research*, *101*, D3, 7461–7475, 1996.

Singh, V. P., Computer models of watershed hydrology, LSU Faculty, Water Resources Publication, 1995.

Sorooshian, S., Q. Y. Duan, and V. K. Gupta, Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting model, *Water Resources Research*, *29*, 4, 1185-1194, 1993.

Tanakamaru, H., Parameter estimation for the tank model using global optimization, *Transactions of the Japanese Society of Irrigation, Drainage and Reclamation Engineering*, *178*, 103–112, 1995.

U.S. Army Corps of Engineers, HEC-1 Flood Hydrograph package, Users and Programmers Manuals, HEC Program 723-X6-L2010, 1973.

U.S. Army Corps of Engineers, Hydrologic Engineering Center, HEC-1, Flood Hydrograph Package, Users Manual, September 1981, Rev. January, 1985.

Van Straten, G., and K. J. Keesman, Uncertainty propagation and speculation in projective forecasts of environmental change: A Lake-Eutrophication example, *Journal of Forecasting*, *10*, 163–190, 1991.

Wang, Q. J., The genetic algorithm and its application to calibrating conceptual rainfall-runoff models, *Water Resources Research*, *27*, 9, 2467–2471, 1991.

Wheater, H. S., S. Tuck, R. C. Ferrier, et al., Hydrological flow paths at the Allt A Mharcaidh Catchment-An analysis of plot and catchment scale observations, *Hydrology Process*, *7*, 4, 359–371, Oct-Dec, 1993.

Woolhiser, D. A., R. E. Smith, and D. C. Goodrich, A kinematic runoff and erosion manual: Documentation and user manual, ARS 77, U.S. Department of Agriculture, 1990.

Yapo, P. O., H. V. Gupta, and S. Sorooshian, Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *Journal of Hydrology*, *181*, 23–48, 1996.

Yapo, P. O., H. V. Gupta, and S. Sorooshian, Multi-objective global optimization for hydrologic models, *Journal of Hydrology*, *204*, 83–97, 1998.

CHAPTER 32

STOCHASTIC CHARACTERISTICS AND MODELING OF HYDROCLIMATIC PROCESSES

JOSÉ D. SALAS AND ROGER A. PIELKE, SR.

1 INTRODUCTION

During 2000, the *Economist* (2001) reported that at least 6 of the top 12 loss of life events and 9 of the top 12 insured property losses were associated with hydrologic events. Late in 1999, an estimated 50,000 lives were lost associated with heavy rain along the northern coast of Venezuela. The quality and quantity of potable water is also important (Pielke and Guenni, 1999). As reported in the *Economist* (May 29, 1999, p. 102), while 90% of the world's population has enough water at present, by 2050 more than 40% of the population is estimated as facing a water shortage. The access to safe water is even more serious. In the same article the *Economist* reports that only about 30% of the rural residents of Brazil currently have access to safe water. Vorosmarty et al. (2000) demonstrate that population growth is the much larger threat to global water resources than any of the current generation projections of future climate. Understanding and quantifying the past, present, and future water availability at the global, regional, and local scales are scientifically, socially, and politically important aspects in balancing water supply and water demand.

Predictability of water resources at any scale requires a good understanding of atmospheric, oceans, and land surface processes and their interactions. In addition, land and oceanic biospheric processes play an important role in the global environment. Figure 1 illustrates the suite of environmental stresses that can threaten water resources. As population increases in a watershed, for example, increased clearing of trees and shrubs, as well as habitation within gulleys and ravines, can increase the vulnerability of the local population to flash flooding. This was a major factor in the

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

large loss of life in the 1999 flood in Venezuela. Assessing the sensitivity of hydrologic processes to landscape change and vegetation dynamics represents one component of Figure 1. To illustrate the procedure to quantitatively assess sensitivity, Figure 2 shows the change in the total model simulated 210-day (during 1989) precipitation over the central United States (Eastman et al., 2001) associated with: (a) the conversion of the current landscape back to its natural form, (b) the radiative effect of doubled atmospheric carbon dioxide, and (c) the biological effect on vegetation of doubled atmospheric carbon dioxide. A coupled atmospheric–vegetation–soil dynamics model was used. The larger scale atmospheric forcing, however, remains identical for the three experiments and is derived from observed National Center for Environmental Prediction analyses (Kalnay et al., 1996).

This analysis shows the surprising result that both landscape change and the biological effect of carbon dioxide can exert a major effect on precipitation. With landscape change, the natural vegetation in the central United States had larger transpiration associated with greater vegetation coverage, particularly tall grass prairie in the eastern portion of the model domain. The increased transpiration cooled the daytime summer atmosphere, thereby preferentially permitting fewer rain showers in the model. Similarly, the enrichment of the atmosphere with carbon dioxide facilitated greater vegetation growth, such that cooling the daytime

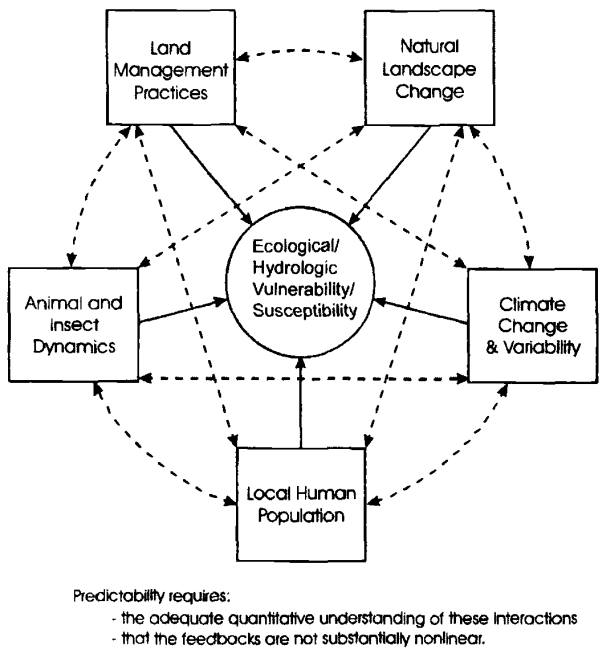


Figure 1 Use of ecological vulnerability/susceptibility in environmental assessment. (*Adapted from Pielke and Guenni, 1999.*)

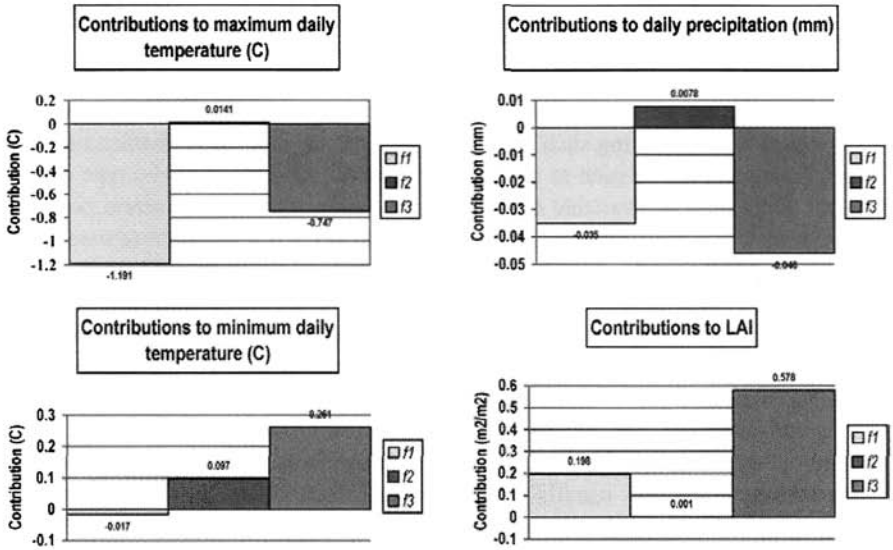


Figure 2 (see color insert) RAMS/GEMTM coupled model results—the seasonal domain-averaged (central Great Plains) for 210 days during the growing season, contributions to maximum daily temperature, minimum daily temperature, precipitation, and leaf area index due to f1 = natural vegetation, f2 = 2XC02 radiation, and f3 = 2xC02 biology. (*Adapted from Eastman et al., 2001*). See ftp site for color image.

atmosphere was also increased for this case. Such experiments illustrate a procedure to assess the sensitivity of hydrologic processes (precipitation in the above example) to environmental change. By assessing the sensitivity of a hydrologic process to the spectrum of environmental stressors, the largest sensitivities can be determined. With this information, social scientists and policy scientists can determine where to most effectively use resources to mitigate or adapt to the environmental threats (Sarewitz et al., 2000).

This brief introduction highlights the importance of the interrelationships and interactions among the various forcing functions of the environment, particularly as they relate to water resources availability and the effect of extremes such as floods and droughts on the environment and on society, and vice versa. Estimating those interactions and effects hinges on the proper characterization of the underlying hydroclimatic processes involved, such as air temperature, precipitation, humidity, snowpack, streamflow, infiltration, soil moisture, sea surface temperature, etc. The rest of this chapter focuses on the characterization and modeling of such processes by using stochastic methods. It is essentially an introduction and overview to two major separate chapters dealing specifically and more in depth with simulation (Salas et al., 2002) and forecasting (Valdes et al., 2002) of hydroclimatic processes particularly precipitation and streamflow.

2 GENERAL CHARACTERISTICS OF HYDROCLIMATIC PROCESSES

Mathematical models are generally used for stochastic simulation and forecasting of hydroclimatic processes. The stochastic characterization of the underlying processes is important in constructing such models. In general, the stochastic characteristics of hydroclimatic processes such as precipitation and runoff depend on the type of data at hand. Data may be available on a continuous time scale or at discrete points in time. For instance, most hydrologic series of practical interest are *discrete time series* defined on hourly, daily, weekly, monthly, bimonthly, quarterly, and annual time intervals. The term *seasonal time series* is often used for series with time intervals that are fractions of a year (usually a month or multiples of a month). Likewise, hourly, daily, weekly, monthly, and seasonal series are often called *periodic-stochastic* series. Hydroclimatic time series may consist of a *single time series* (*univariate series*) or *multiple time series* (*multivariate series*).

Hydroclimatic time series are generally *autocorrelated*. Autocorrelation in some series such as streamflow usually arises from the effect of surface, soil, and groundwater storages that cause the water to remain in the system through subsequent time periods (Salas, 1993). For instance, basins with significant surface storage in the form of lakes, swamps, or glaciers, produce streamflow series that are autocorrelated. Likewise, subsurface storage, especially groundwater storage produces significant autocorrelation in the streamflow series derived from groundwater outflow. Conversely, annual precipitation and annual maximum flows (flood peaks) are usually uncorrelated. Sometimes significant autocorrelation may be the result of trends and/or shifts in the series (Salas and Boes, 1980; Eltahir, 1989). In addition, multiple hydroclimatic series may be *cross-correlated*. For example, the precipitation series at two nearby sites, or the streamflow series of two nearby gaging stations in a river basin are expected to be cross-correlated because the sites are subject to similar climatic and hydrologic events. As the sites considered become farther apart, their cross-correlation decreases. However, because of the effect of some large-scale atmospheric-oceanic phenomena such as El Niño Southern Oscillation (ENSO), significant cross-correlation between sea surface temperature (SST) and streamflow between sites thousands of miles apart can be found (Eltahir, 1996). Furthermore, one would expect a significant cross-correlation between a streamflow time series and the corresponding areal average precipitation series over the same basin.

Hydroclimatic time series are *intermittent* when the variable under consideration takes on nonzero and zero values throughout the length of the record. For instance, the precipitation that is observed in a recording rain gage is an intermittent time series. Likewise, hourly, daily, and weekly rainfall are typically intermittent time series, while monthly and annual rainfall are usually nonintermittent. However, in semiarid and arid regions even monthly and annual precipitation and monthly and annual runoff may be intermittent as well.

Traditionally, certain annual hydroclimatic series have been considered to be *stationary*, although this assumption may be incorrect as a result of large-scale climatic variability, natural disruptions such as a volcanic eruption, and anthropogenic changes such as the effect of reservoir construction on downstream flow, and

the effect of landscape changes on some components of the hydrologic cycle. On the other hand, hydroclimatic series defined at time intervals smaller than a year, such as months, generally exhibit distinct *seasonal* (*periodic*) patterns due to the annual revolution of Earth around the sun, which produces the annual cycle in most hydroclimatic processes. Some series of interest to hydrology and water resources, such as daily urban water use, may also exhibit a *weekly pattern* due to variations of demands within a week. Likewise, hourly time series may have a distinct *diurnal pattern* due to the variations of demands within a day. Summer hourly rainfall series or certain water quality constituents related to temperature may also exhibit distinct diurnal patterns due to the daily rotation of Earth that causes variations of net radiation within the day (Obeysekera et al., 1987; Katz and Parlange, 1995). Seasonal patterns of hydroclimatic series translate into statistical characteristics that vary within the year (or within a week or a day as the case may be) such as seasonal or periodic variations in the mean, variance, covariance, and skewness. Removing the seasonality in the mean and in the variance has been generally accomplished by the so-called *seasonal standardization*. This procedure is often referred to in the literature as *deseasonalization*. Unfortunately, this term is a misnomer since it may imply that the residual series is free of seasonality. However, seasonality may still be present in the covariance structure as is generally the case for seasonal streamflow series (Salas, 1993).

Hydroclimatic time series may exhibit trends, shifts or jumps, seasonality, autocorrelation, and non-normality. These attributes of hydroclimatic time series are referred to as *components* (Salas, 1993). In general, natural and human-induced factors may produce gradual and instantaneous trends and shifts (jumps) in hydroclimatic series. For example, a large forest fire in a river basin can immediately affect the runoff, producing a shift in the runoff series, whereas a gradual killing of a forest (e.g., by an insect infestation that takes years for its population to build up) can result in gradual changes or trends in the runoff series. A large volcanic explosion such as the one at Mount St. Helens in 1980 or a large landslide can produce sudden changes in the sediment transport series of a stream. Trends in non-point-source water quality series may be the result of long-term changes in agricultural practices and agricultural land development. Likewise, shifts in certain water quality constituents may be caused by agricultural activities such as sudden changes in the use of certain types of pesticides. Changes in land use and the development of reservoirs and diversion structures may also cause trends and shifts in streamflow series. The current concern about global warming and large-scale climatic variability, such as shifts in the intertropical convergence zone (ICZ) and the effects of large-scale oscillations such as ENSO and the Pacific Decadal Oscillation (PDO), is making hydroclimatologists more aware of the occurrence of trends and shifts in hydroclimatic time series. Figure 3 illustrates the observed swings or shifts of the time series of standardized deviations of annual rainfall for Central and West Sahel areas during the period 1950–1998 (Landsea et al., 1999). Concerns regarding the effects of such types of sudden shifts observed in some hydroclimatic time series on water resources, the environment, and society have been expressed and documented in the literature (e.g., Kerr, 1992; Taylor, 1999). Statistical techniques are available for detecting,

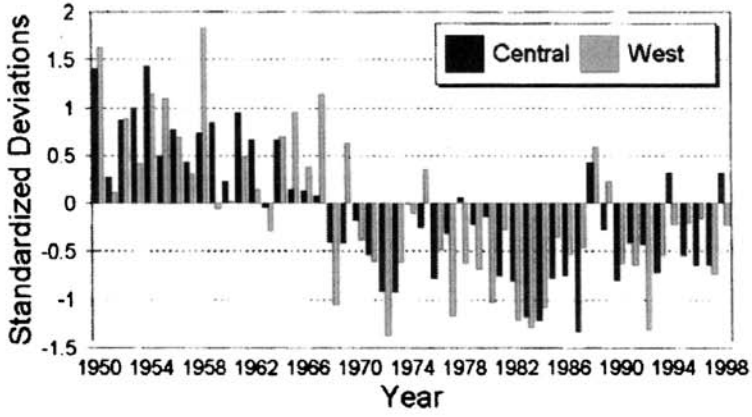


Figure 3 (see color insert) Swings or shifts of the time series of standardized deviations of annual rainfall for Central and West Sahel areas during the period 1950–1998. (*After Landsea et al., 1999.*) See ftp site for color image.

modeling, and removing trends and shifts from hydroclimatic times series (Helsel and Hirsch, 1992; Salas, 1993; Hipel and McLeod, 1994).

3 STOCHASTIC ANALYSIS AND PROPERTIES OF HYDROCLIMATIC TIME SERIES

Overall Statistical Characteristics

The most commonly used statistical properties for analyzing stationary or nonstationary hydroclimatic time series are the sample mean $\bar{y}$, variance s^2, coefficient of variation cv, skewness coefficient g, lag-k autocorrelation coefficient r_k, and the spectrum $g(f)$. Coefficients of variation of annual flows are typically smaller than one, although they may be close to one or greater in streams in arid and semiarid regions. The coefficients of skewness g of annual flows are typically greater than zero. In some streams, small values of g are found suggesting that annual flows are approximately normally distributed. On the other hand, in some streams of arid and semiarid regions, g can be greater than one.

The lag-k autocorrelation coefficient r_k may be determined as

$$r_k = \frac{c_k}{c_0} \qquad k = 0, 1, 2, \ldots \tag{1a}$$

$$c_k = \frac{1}{N} \sum_{t=1}^{N-k} (y_{t+k} - \bar{y})(y_t - \bar{y}) \tag{1b}$$

where N is the sample size and k is the time lag. The plot of r_k versus k, i.e., the *correlogram*, may give an idea of the degree of persistence of the underlying time series, and it may be useful for choosing the type of stochastic model that may represent the series. When the correlogram decays rapidly to zero after a few lags, it may be an indication of small *persistence* or *short memory* in the series, while a slow decay of the correlogram is an indication of large persistence or *long memory*. The lag-one serial correlation coefficient r_1 is a simple measure of the degree of time dependence of a series. Generally, r_1 for annual flows is small but positive, although negative r_1's may occur because of sample variability. Large values of r_1 for annual flows can be found for a number of reasons including the effect of natural or man-made surface storage such as lakes, reservoirs, or glaciers, the effect of slow groundwater storage response, and the effect of nonstationarity. The estimators s^2, g, and r_k are biased (downward relative to the corresponding population statistics). Corrections for bias for these estimators have been suggested (Bobee and Robitaille, 1975; Yevjevich, 1972a; Fernandez and Salas, 1990).

In addition, the sample spectrum is another way of studying the variability of hydroclimatic series in the frequency domain (Yevjevich, 1972b). The sample spectrum $g(f_j)$ may be determined as

$$g(f_j) = 2\left[1 + 2\sum_{k=1}^{m} D_k r_k \cos(2\pi f_j k)\right] \qquad f_j = \frac{j}{2m} \quad j = 0, 1, 2, \ldots, m \tag{2}$$

where D_k is a smoothing function and m is the maximum number of lags considered. Figure 4 illustrates the autocorrelation function and the spectrum obtained for the time series of annual PDO indices for the period 1900–1999. The time series shows evidence of low-frequency components, which are manifested in a slow decaying and pseudoperiodic correlogram and a spectrum with visible high values at frequencies near 0.02 and 0.18 cycles per year.

When analyzing several time series jointly, cross-correlations may be important. The cross-correlation coefficient between series $y_t^{(i)}$ and $y_t^{(j)}$, $t = 1, \ldots, N$ for stations i and j, is determined as

$$r_k^{ij} = \frac{c_k^{ij}}{(c_0^{ii} c_0^{jj})^{1/2}} \qquad k = \cdots -2, -1, 0, 1, 2, \ldots \tag{3a}$$

$$c_k^{ij} = \frac{1}{N}\sum_{t=1}^{N-k} (y_{t+k}^{(i)} - \bar{y}^{(i)})(y_t^{(j)} - \bar{y}^{(j)}) \tag{3b}$$

The plot of r_k^{ij} vs. k is the *cross-correlogram*. For n time series, the values of r_k^{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, n$ are elements of the lag-k cross-correlation $n \times n$ matrix $\hat{M}_k$. Figure 5 is a graphical display of the lag-zero cross-correlation matrix obtained for the annual streamflows of 29 stations in the Colorado River system. For reference, station 1 is one of the farthest upstream site while station 29 is the farthest downstream site. The cross-correlation between stations 1 and 29 is large (of the order of 0.9) while the cross-correlation between stations 1 and 27 is small (the

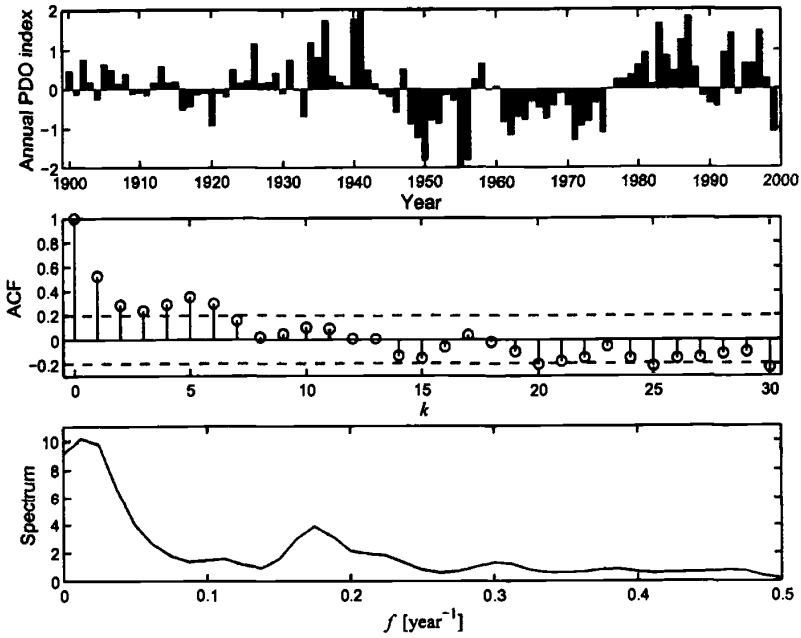


Figure 4 Autocorrelation function and spectrum obtained for the time series of annual PDO indices for the period 1900–1999. The time series shows evidence of low-frequency components. (*From Oli Sveinsson, Ph.D. candidate, CSU.*)

reason being that station 27 is a small tributary of the Colorado River and is very far from station 1).

In modeling hydroclimatic time series such as streamflow for simulation studies of reservoir systems, storage-related stochastic properties such as the range of cumulative departures R_n^*, the rescaled range R_n^{**}, and the Hurst slope K may be particularly important. They have been widely used in the literature as measures of long-term dependence and for comparing alternative models of hydrologic series (Hurst, 1951; Wallis and O'Connell, 1973; Hipel and McLeod, 1994). In particular, Hurst (1951) showed that for a large number of geophysical time series such as streamflow, precipitation, temperature, and tree-ring series, the mean rescaled range $\bar{R}_n^{**}$ (n = sample size) is proportional to n^h with $h > \frac{1}{2}$. The values of h obtained for different series gave a mean of about 0.73 and a standard deviation of 0.09. Theoretical results for normal independent processes and for autoregressive processes (Mandelbrott and Van Ness, 1968) indicated that asymptotically $h = \frac{1}{2}$. The discrepancy between theoretical results stating that $h = \frac{1}{2}$ and Hurst empirical findings suggesting that $h > \frac{1}{2}$ has become known as the *Hurst phenomenon*. However, the estimates of h are transient, meaning they depend on n and as $n \to \infty$, they generally converge to a limiting value, equal to $\frac{1}{2}$ for many time series models (Salas et al.,

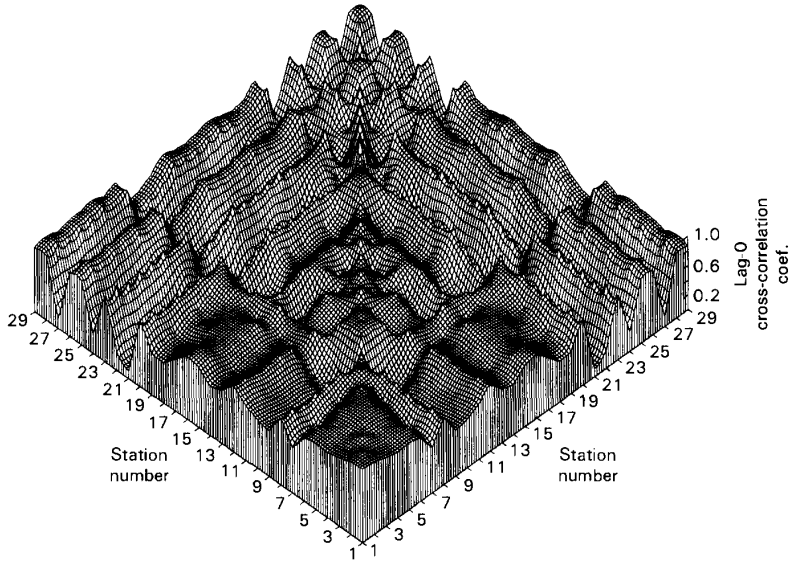


Figure 5 Lag-zero cross-correlation matrix obtained for the annual streamflows of 29 stations in the Colorado River system. For reference station 1 is the furthest upstream site while station 29 is the furthest downstream site.

1979). One interpretation of the Hurst phenomenon has been to associate $h = \frac{1}{2}$ with short memory models possessing short-term dependence structure, and $h > \frac{1}{2}$ with long memory models possessing long-term dependence. A number of models, including the autoregressive moving average (ARMA) processes, can have long-term dependence structure, yet asymptotically they give $h = \frac{1}{2}$. Furthermore, a stationary model with long-term dependence and $h > \frac{1}{2}$ is the fractional ARMA (FARMA) model (refer to Section 4 for definitions of ARMA and FARMA processes). Estimates of h can be useful for comparing the performance of alternative modeling strategies and estimation procedures. Statistical tests to determine whether a given time series exhibits the Hurst effect are also available (Mesa and Poveda, 1993).

Furthermore, drought-related stochastic properties are also important in modeling some hydroclimatic time series such as precipitation and streamflow. Consider a hydrologic time series y_t, $t = 1, \ldots, N$, and a *demand level* d (*crossing level*). Assume that y_t is an annual series and d is a constant (e.g., $d = \alpha\bar{y}$ and $0 < \alpha \leq 1$). A deficit at any given time t occurs when $y_t < d$. A consecutive sequence of deficits (until $y_t > d$ again) may be called a drought, and such a drought can be characterized by its duration L, its magnitude M, and its intensity $I = M/L$ (Yevjevich, 1967). Because a number of droughts can occur in a given hydrologic sample, the maximum drought duration, magnitude, and intensity (in a given

sample) have been indicators of the so-called *critical drought* and have been widely used in water resources studies.

Periodic (Seasonal) Statistical Properties

While overall stochastic properties of hydroclimatic time series, such as those previously defined above, may be determined from either annual series or for seasonal series as a whole, specific seasonal (periodic) properties may provide a better picture of the stochastic characteristics of certain hydroclimatic time series that are defined at time intervals smaller than a year such as monthly streamflow data. Let the seasonal time series be represented by $y_{v,\tau}$, $v = 1, \ldots, N$; $\tau = 1, \ldots, \omega$ in which v is the year, τ is the season, N is the number of years of record, and ω is the number of seasons per year (e.g., $\omega = 12$ for monthly data). Then, for each season τ one can determine a number of statistics such as the seasonal mean $\bar{y}_\tau$, variance s_τ^2, coefficient of variation cv_τ, and skewness coefficient g_τ. Furthermore, the season-to-season correlation coefficient $r_{k,\tau}$ may be estimated by

$$r_{k,\tau} = \frac{c_{k,\tau}}{(c_{0,\tau-k}c_{0,\tau})^{1/2}} \qquad k = 0, 1, 2, \ldots; \quad \tau = 1, \ldots, \omega \tag{4a}$$

$$c_{k,\tau} = \frac{1}{N}\sum_{v=1}^{N}(y_{v,\tau} - \bar{y}_\tau)(y_{v,\tau-k} - \bar{y}_{\tau-k}) \tag{4b}$$

For instance, for monthly streamflows $r_{1,4}$ represents the correlation between the flows of the fourth month with those of the third month. Likewise, for multiple seasonal time series, the lag-k seasonal cross-correlation coefficient $r_{k,\tau}^{ij}$ between the seasonal time series $y_{v,\tau}^{(i)}$ and $y_{v,\tau-k}^{(j)}$ for sites i and j, can be determined.

The statistics $\bar{y}_\tau$, s_τ, g_τ, and $r_{k,\tau}$ may be plotted versus time $\tau = 1, \ldots, \omega$ to observe whether they exhibit a seasonal pattern. Fitting these statistics by Fourier series is especially effective with weekly and daily data (Salas et al., 1980). Generally, for seasonal streamflow series $\bar{y}_\tau > s_\tau$ although for some streams $\bar{y}_\tau$ may be smaller than s_τ especially during the "low-flow" season. Furthermore, for intermittent streamflow series generally the mean is smaller than the standard deviation, i.e., $\bar{y}_\tau < s_\tau$ throughout the year. Likewise, values of the skewness coefficient g_τ for the dry season are generally larger than those for the wet season indicating that data in the dry season depart more from normality than data in the wet season. Values of the skewness for intermittent hydrologic series are usually larger than skewness for similar nonintermittent series. Seasonal correlations $r_{k,\tau}$ for streamflow during the dry season are generally larger than those for the wet season, and they are significantly different than zero for most of the months. Figure 6 displays $r_{1,\tau}$, i.e., the lag-1 month-to-month correlations, for the monthly streamflows of the referred 29 stations of the Colorado River system. It may be observed that the correlations vary with the month, and with a few exceptions the correlation pattern for the entire system is similar. On the other hand, seasonal correlations for monthly precipitation are generally low or not significantly different from zero for most of the months (Roesner and

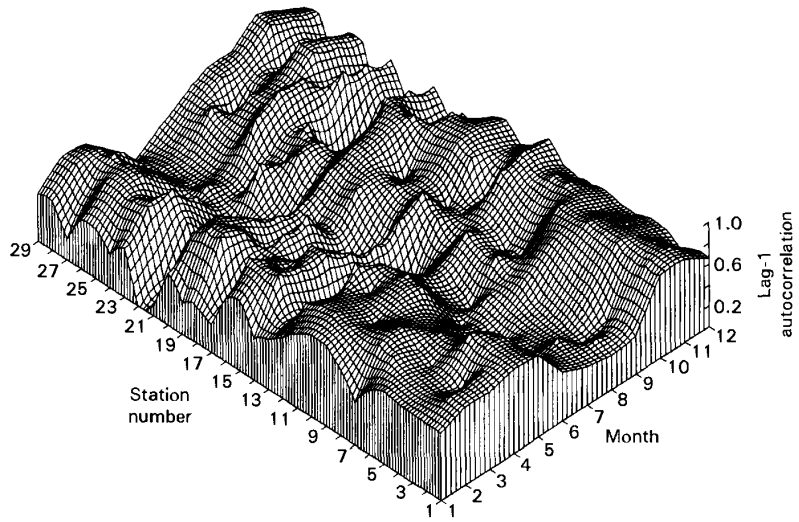


Figure 6 The lag-1 month-to-month correlations, i.e., $r_{1,\tau}$ for the monthly streamflows of 29 stations of the Colorado River system.

Yevjevich, 1966), while for weekly, daily, and hourly precipitation they are generally significant and greater than zero.

Complex, long-term dependence (long memory) of seasonal flows may be evident when the correlations $r_{k,\tau}$ are significant and decay slowly as k increases beyond ω seasons (beyond a year). These correlations are usually small or not significant for many streams, but in river systems such as the Nile River such seasonal correlations may persist for several years. Rivers that exhibit long-term correlation in seasonal flows will exhibit also long-term autocorrelation in the annual flows. In addition, some streamflow hydrographs such as daily and weekly hydrographs may possess directionality (nonreversibility), which means that some of their statistical properties change when direction of time is reversed. This is evident from the typical form of hydrographs in which the rising limb is shorter than the recession limb. In these cases, it is desirable that the mathematical models have such directionality attribute (Fernandez and Salas, 1986).

4 STOCHASTIC MODELS AND MODELING TECHNIQUES

A number of stochastic models and modeling schemes have been developed for simulation and forecasting of hydroclimatic processes. Some of the models are conceptually (physically) based, some others are empirical or transformed or adapted from existing models developed in other fields, while some others have arisen specifically to address some particular features of the process under consideration.

In general models for continuous time processes and models for short time scales such as hourly are more complex than models for larger time scales. Also some of the models have been developed specifically for precipitation while some others are for streamflow. Yet many of them are useful for both and for many other hydroclimatic processes. We will illustrate here as a matter of introduction and subsequent reference, the family of autoregressive and moving average (ARMA) models and extensions and modifications thereof. These models have become quite popular for both simulation and forecasting of many hydroclimatic processes. However, many other stochastic models have been developed, some of them quite different than ARMA models, aimed at the specific process under consideration or the particular features (of the underlying process) one tries to address. For example, for intermittent processes such as daily rainfall, Markov chains and the discrete counterpart of ARMA models, i.e., discrete ARMA (DARMA), are available (e.g., Chang et al., 1984; Guttorp, 1995). Likewise, models with infinite memory such as the Fractional Gaussian noise (e.g., Mandelbrot and Van Ness, 1968) and shifting level models that are capable of simulating sudden shifts (e.g., Salas and Boes, 1980) are available.

Stochastic Models

Stationary Models. The family of ARMA models has been widely used for modeling hydroclimatic processes at various time scales. The ARMA(p, q) model is defined as (Brockwell and Davis, 1991)

$$y_t = \mu + \sum_{j=1}^{p} \phi_j (y_{t-j} - \mu) + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{5a}$$

$$\phi(B)(y_t - \mu) = \theta(B)\varepsilon_t \tag{5b}$$

where μ, the ϕ's, the θ's, and $\sigma^2(\varepsilon)$ are parameters of the model, p is the order of the autoregressive terms, q is the order of the moving average terms, $B^i z_t = z_{t-i}$, and

$$\phi(B) = 1 - \phi_1 B^1 - \phi_2 B^2 - \cdots - \phi_p B^p \tag{6a}$$

$$\theta(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q \tag{6b}$$

Particular models derived from (5) are the ARMA$(p, 0)$ or AR(p) and the ARMA$(0, q)$ or MA(q) models. In addition, the fractional autoregressive moving average FARMA(p, d, q) model is defined as (Hosking, 1981; Montanari et al., 1997)

$$\phi(B)(1 - B)^d (y_t - \mu) = \theta(B)\varepsilon_t \qquad -0.5 < d < 0.5 \tag{7}$$

This model is capable of representing long-term dependence. The foregoing ARMA and FARMA models are stationary, hence their applications to modeling

hydroclimatic time series require that the underlying data be stationary or be converted to stationary by some appropriate transformation.

These models have been generally applied to annual hydroclimatic data. Sometimes they have been applied to seasonal data after seasonal standardization. Likewise, they have been applied to daily data either after seasonal standardization or by separating the year into several seasons and applying different models to the daily series in each season. For example, Parlange et al. (1992) applied physically based concepts to the daily variations of soil moisture and found that it can be described by an AR(1) process. Properties of AR and ARMA models, such as the autocorrelation function, variance, and spectrum, and hydrologic applications may be found in Salas et al. (1980), Loucks et al. (1981), Bras and Rodriguez-Iturbe (1985), Salas (1993), and Hipel and McLeod (1994). Also Chu and Katz (1985) fitted AR and ARMA models to seasonal and monthly Southern Oscillation Index (SOI). Before fitting the models, the annual cycle was removed from the data in order to make them stationary. Xu and Storch (1990) used Principal Oscillation Pattern (POP) analysis to model monthly SOI data. They concluded that their POP scheme was superior over the ARMA scheme. Furthermore, Chu et al. (1995) applied a bivariate AR model for modeling jointly the seasonal SOI and a precipitation index in Florida. The fitted bivariate AR model was then used to forecast precipitation.

Periodic Models. A number of periodic and other nonstationary models such as the family of PARMA, ARIMA, and multiplicative PARMA models has been suggested in the literature for modeling seasonal hydroclimatic processes such as seasonal precipitation and streamflow series (Salas et al., 1980; Loucks et al., 1981; Salas, 1993; Hipel and McLeod, 1994). In particular, the PARMA(p,q) model is defined as

$$y_{v,\tau} = \mu_\tau + \sum_{j=1}^{p} \phi_{j,\tau}(y_{v,\tau-j} - \mu_{\tau-j}) + \varepsilon_{v,\tau} - \sum_{j=1}^{q} \theta_{j,\tau}\varepsilon_{v,\tau-j} \tag{8a}$$

$$\phi_\tau(B)(y_{v,\tau} - \mu_\tau) = \theta_\tau(B)\varepsilon_{v,\tau} \tag{8b}$$

where

$$\phi_\tau(B) = 1 - \phi_{1,\tau}B^1 - \phi_{2,\tau}B^2 - \cdots - \phi_{p,\tau}B^p \tag{9a}$$

$$\theta_\tau(B) = 1 - \theta_{1,\tau}B^1 - \theta_{2,\tau}B^2 - \cdots - \theta_{q,\tau}B^q \tag{9b}$$

and $B^i z_{v,\tau} = z_{v,\tau-i}$. When $q = 0$, the foregoing model becomes the well-known PARMA(p, 0) or PAR(p). More specifically the PAR(1) model (also known as the Thomas–Fiering model) is likely one of the most widely used models in hydrology. In general low-order PARMA models have become popular for modeling seasonal hydroclimatic processes. Physically based or conceptual arguments of the underlying hydrologic cycle of a watershed or river basin justify the applicability of these models. For instance, Salas and Obeysekera (1992) showed that assuming that the precipitation input is an uncorrelated periodic-stochastic process and under some

linear reservoir considerations for the groundwater storage, the stochastic model for seasonal streamflow becomes a PARMA(1,1) process. Chu et al. (1995) analyzed time series of seasonal and monthly SOI and fitted AR and ARMA models after the annual cycle was removed from the data. They also used ARMA models with seasonally varying coefficients.

In addition, ARIMA(p, d, q), multiplicative ARMA, and multiplicative ARIMA models have been applied for forecasting hydroclimatic processes (e.g., Salas et al., 1980; Hipel and McLeod, 1994), sampling groundwater levels (Ahn and Salas, 1997), and for detection and estimation of trends in climatological time series (e.g., Visser and Molenaar, 1995; Zheng and Basher, 1999). Furthermore, simulation of complex processes such as the Nile River monthly flows has been accomplished with multiplicative PARMA models (Salas et al., 1995). Also Lund et al. (1995) provide a general overview of the analysis and modeling of climatological time series having periodic correlation structure. They suggest a test for detection of periodic correlation and applied PARMA models for modeling such series. While the referred stationary and nonstationary models [e.g., models (5) and (8), respectively] are written for single-site or univariate series, their multisite or multivariate counterparts are also available (e.g. Salas, 1993; Hipel and McLeod, 1994).

Stochastic Models for Forecasting

A number of stochastic models have been widely applied for forecasting hydroclimatic processes such as precipitation and streamflow. Many of such models fall in the family of transfer function models. The general transfer function noise (GTFN) model may be written as

$$\gamma(B)(y_t - \mu_y) = \frac{\omega(B)}{\delta(B)}(x_{t-\tau} - \mu_x) + \frac{\theta(B)}{\phi(B)}\varepsilon_t \tag{10}$$

where γ(B), ω(B), δ(B), θ(B), and ϕ(B) are polynomials in B of different orders [similar to those defined in Eq. (6)], x_t is the exogenous variable such as precipitation or ENSO index, the μ's represent the means, τ is the time delay, and ε is the noise term. Some special cases (models) such as the ARMA, ARMAX, unit hydrograph type, multiple linear regression, and the Box–Jenkins transfer function noise models can be derived or simplified from (10). Equation (10) assumes single-site variables, but they are applicable to multisite variables if the variables are vectors and the parameters are matrices. Applications of many of these models can be found in Hipel and McLeod (1994). In addition, forecasting equations based on ARMA, ARMAX, and GTFN models can be written in sequential and recursive forms (e.g., using a Kalman filter). Furthermore, artificial neural networks (ANN) have emerged in the last decade as a useful technique for many modeling applications including forecasting (e.g., Hsu et al., 1995; Govindaraju and Rao, 2000). The application of many of these models, estimation procedures, and ANN algorithms for forecasting precipitation and streamflow are described in some detail in Valdes et al. (2001).

Modeling Schemes

Some specific models have been developed in the hydrologic field to address some unique features related to hydrologic and water resources problems. An example is the so-called *disaggregation* models (e.g., Valencia and Schaake, 1973). The failure of some traditional models such as the PAR(1) model to reproduce annual statistics (or upscale statistics) led to the development of disaggregation techniques. While the main intent of such disaggregation models has been to enable one to generate hydrologic sequences that can reproduce statistics at the annual and seasonal time scales, it has brought a major dimension into the capability of modeling complex hydrologic processes and complex hydrologic systems. Complex systems involve several sites, and the temporal and spatial mass balance requirements, often require the use of *modeling schemes* that may consist of an array of single-site, multisite, and temporal and spatial disaggregation models. While this requirement has been more evident in models constructed for simulation, the same is true for forecasting complex hydrologic systems. Furthermore, in order to facilitate the practical application of stochastic models for simulation of hydrological processes, software packages such as SPIGOT (Grygier and Stedinger, 1990) and SAMS (Salas et al., 2000) have been developed. Still, actual applications of such packages in real-world systems, especially for simulating complex hydrologic processes and complex water resources systems such as the Great Lakes system in North America or the Nile River system in Africa, may not be a straightforward application. Thus adjustments, modifications, additions, etc., may have to be made before a satisfactory or acceptable solution to the problem is attained.

Stochastic Modeling

Stochastic modeling of hydroclimatic processes may involve four major steps: model identification, parameter estimation, model testing, and model verification. By model identification is meant determining a specific model structure and the model order; for example, determining that the model for annual streamflow series is an ARMA(1,1) or determining that the model for daily rainfall is a simple Markov chain. Generally models that belong to the family of ARMA, ARIMA, and transfer function models are amenable for certain identification procedures based on autocorrelation, partial correlation, and cross-correlation analysis (Brockwell and Davis, 1991; Hipel and McLeod, 1994). However, model identification techniques are not available for some models or they are too complex, so instead a model of a certain type and order is applied to the particular hydroclimatic series at hand and its performance is judged by testing and verification. Some hydroclimatic processes such as streamflow and soil moisture have been identified using physically based concepts and arguments (e.g. Salas and Smith, 1981; Parlange et al., 1992; Salas and Obeysekera, 1992).

Once a model is identified, its parameters may be estimated by a number of techniques such as the method of moments, least squares, and maximum likelihood, depending on the particular model and data at hand. Typical method of moments

estimation procedures involve matching historical and population (model) first- and second-order statistics, although in some cases some other properties such as skewness and storage and drought related statistics have been used (e.g., Salas et al., 1980; Hipel and McLeod, 1994). In addition, recursive parameter estimation methods and filtering techniques have been used particularly for forecasting problems (e.g., Bras and Rodriguez-Iturbe, 1985). Furthermore, modeling of hydroclimatic time series either for simulation or forecasting generally requires that the underlying series be transformed to approximately normally distributed series (e.g., Salas, 1993; Hipel and McLeod, 1994). Thus parameter estimation is usually made in the transformed domain. Model testing procedures have been well developed for models within the ARMA, ARIMA, ARMAX, and transfer function type of models (e.g., Brockwell and Davis, 1991). Likewise, testing procedures are available for PARMA models (e.g., Salas et al., 1980; Salas, 1993; Hipel and McLeod, 1994). The tests usually involve diagnostic checks to verify whether the model residuals comply with the underlying assumptions of independence and normality (of the residuals). Since many models may comply with such requirements, a model selection criteria based on the Akaike Information Criteria (AIC) is available to discriminate and find a parsimonious model (Brockwell and Davis, 1991). On the other hand, model testing for some other models cannot be done based on analysis of residuals; so instead model testing is based on data generation experiments. In addition, model verification is usually needed beyond testing residuals depending on whether the modeling exercise is geared to simulation or forecasting. For instance, for simulation (data generation) one may like to test whether the model is capable of generating sequences that reproduce a number of storage and drought related historical characteristics. This is usually accomplished by Monte Carlo experiments. On the other hand, model verification for forecasting may involve examining whether the model is capable of estimating the hydrologic process under consideration one or more lead times in advance within a specified error criteria. This may be done by split sampling estimation and testing.

REFERENCES

Ahn, H., and J. D. Salas, Groundwater head sampling based on stochastic analysis, *Water Resour. Res.*, *33*(12), 2769–2780, 1997.

Bobee, B., and R. Robitaille, Correction of bias in the estimation of the coefficient of skewness, *Water Resour. Res.*, *11*(6), 851–854, 1975.

Bras, R. L., and Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, MA, 1985.

Brockwell, P. J., and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed., Springer-Verlag, New York, 1991.

Chang, T. J., M. L. Kavvas, and J. W. Delleur, Daily precipitation modeling by discrete autoregressive moving average processes, *Water Resour. Res.*, *20*(5), 565–580, 1984.

Chu, P. S., and R. W. Katz, Modeling and forecasting the Southern Oscillation: A time-domain approach, *Monthly Weather Rev.*, *113*, 1876–1888, 1985.

Chu, P. S., R. W. Katz, and P. Ding, Modeling and forecasting seasonal precipitation in Florida: A vector time-domain approach, *Int. J. Climatol.*, *15*, 53–64, 1995.

Eastman, J. L., M. B. Coughenour, and R. A. Pielke, The effects of CO_2 and landscape change using a coupled plant and meteorological model, *Global Change Biol.*, 7, 797–815, 2001.

Economist, Catastrophes, March 31, 106, 2001.

Eltahir, E. A. B., A feedback mechanism in annual rainfall in Central Sudan, *J. Hydrol.*, 110, 323–334, 1989.

Eltahir, E. A. B., El Niño and the natural variability in the flow of the Nile River, *Water Resour. Res.*, *32*(1), 131–137, 1996.

Fernandez, B., and J. D. Salas, Periodic gamma autoregressive processes for operational hydrology, *Water Resour. Res.*, *22*(10), 1385–1396, 1986.

Fernandez, B., and J. D. Salas, Gamma-autoregressive models for streamflow simulation, *J. Hydrol. Eng. ASCE*, *116*(11), 1403–1414, 1990.

Govindaraju, R., and A. R. Rao (Eds.), *Artificial Neural Networks in Hydrology*, Kluwer Academic, London, 2000.

Grygier, J. C., and J. R. Stedinger, *SPIGOT, A Synthetic Streamflow Generation Software Package, Technical Description*, Version 2.5, *Cornell University*, School of Civil and Environmental Engineering, Ithaca, NY, 1990.

Guttorp, P., *Stochastic Modeling of Scientific Data*, Chapman & Hall, London, 1995.

Helsel, D. R., and R. M. Hirsch, *Statistical Methods in Water Resources*, Studies in Environmental Science 49, Elsevier, Amsterdam, 1992.

Hipel, K. W., and A. I. McLeod, *Time Series Modeling of Water Resources and Environmental Systems*, Elsevier, Amsterdam, 1994.

Hosking, J. R. M., Fractional differencing, *Biometrika*, *68*, 165–176, 1981.

Hsu, K., H. V. Gupta, and S. Sorooshian, Artificial neural network modeling of the rainfall-runoff process, *Water Resour. Res.*, *31*(10), 2517–2530, 1995.

Hurst, H. E., Long-term storage capacity of reservoirs, *Trans. ASCE*, *116*, 770–799, 1951.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471, 1996.

Katz, R. W., and M. B. Parlange, Generalizations of chain-dependent processes: Application to hourly precipitation, *Water Resour. Res.*, *31*(5), 1331–1341, 1995.

Kerr, R. A. Unmasking a shifty climate system, *Research News*, *255*, 1508–1510, 1992.

Landsea, C. W., W. M. Gray, P. W. Mielke Jr., K. J. Berry, and R. K. Taft, June to September rainfall in North Africa: A seasonal forecast for 1999, Atmospheric Science Department, Colorado State University, Fort Collins, CO, available on-line, http://typhoon.atmos.colostate.edu/forecasts/1999/sahel_jun99/,1999.

Loucks, D. P., J. R. Stedinger, and D. Haith, *Water Resources Systems Planning and Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

Lund, R., H. Hurd, P. Bloomfield, and R. Smith, Climatological time series with periodic correlation, *J. Climate*, *8*, 2787–2809, 1995.

Mandelbrot, B. B., and J. W. Van Ness, Fractional Brownian motions, fractional noises and applications, *SIAM Rev.*, *10*(4), 422–437, 1968.

Mesa, O. J., and G. Poveda, The Hurst effect: The scale of fluctuation approach, *Water Resour. Res.*, *29*(12), 3995–4002, 1993.

Montanari, A., R. Rosso, and M. S. Taqqu, Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation, *Water Resour. Res.*, *33*(5), 1035–1044, 1997.

Obeysekera, J. T. B., G. Tabios, and J. D. Salas, On parameter estimation of temporal rainfall models, *Water Resour. Res.*, *23*(10), 1837–1850, 1987.

Parlange, M. B., G. G. Katul, R. H. Cuenca, M., L. Kavas, D. R. Nielsen, and M. Mata, Physical basis for a time series model of soil water content, *Water Resour. Res.*, *28*(9), 2437–2446, 1992.

Pielke, Sr., R. A., and L. Guenni, Vulnerability assessment of water resources to changing environmental conditions, *IGBP Newsl.*, *39*, 21–23, 1999.

Roesner, L. A., and V. Yevjevich, Mathematical models for time series of monthly precipitation and monthly runoff, in *Hydrology Papers*, No. 15, Colorado State University, Ft. Collins, CO, 1966.

Salas, J. D., *Analysis and Modeling of Hydrologic Time Series*, in D. R. Maidement (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993, Chapter 19.

Salas, J. D., and D. C. Boes, Shifting level modeling of hydrologic series, *Adv. Water Resour.*, *3*, 59–63, 1980.

Salas, J. D., and J. T. B. Obeysekera, Conceptual basis of seasonal streamflow time series models, *ASCE J. Hydraul. Eng.*, *118*(8), 1186–1194, 1992.

Salas, J. D., and R. A. Smith, Physical bases of stochastic models of annual flows, *Water Resour. Res.*, *17*, 428–430, 1981.

Salas, J. D., D. C. Boes, V. Yevjevich, and G. G. S. Pegram, Hurst phenomenon as a pre-asymptotic behavior, *J. Hydrol.*, *44*(1), 1–15, 1979.

Salas, J. D., J. R. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, Water Resources Publications, Littleton, CO, 1980.

Salas, J. D., N. Saada, and C. H. Chung, Stochastic modeling and simulatio of the Nile River system monthly flows, Colorado State University, Engineering Research Center, Comp. Hydrol. Lab., Technical Report No. 5, Fort Collins, CO, 1995.

Salas, J. D., N. Saada, C. H. Chung, W. L. Lane, and D. K. Frevert, Stochastic analysis, modeling, and simulation (SAMS), Version 2000, User's manual, Colorado State University, Engineering Research Center, Comp. Hydrol. Lab., Technical Report No. 10, Fort Collins, CO, 2000.

Salas, J. D., J. A. Ramirez, P. Burlando, and R. Pielke, Sr., Stochastic simulation of precipitation and streamflow processes, in T. D. Potter and B. Colman (Eds.), *Handbook of Weather, Climate, and Water*, Chapter 33 John Wiley & Sons, Inc., New York, 2002.

Sarewitz, D., R. A. Pielke, Jr., and R. Byerly (Eds.), *Prediction: Science Decision Making and the Future of Nature*, Island Press, Covelo, CA, 2000.

Taylor, K. Rapid climate change, *Am. Sci.*, *87*, 320–326, 1999.

Valdes, J. B., P. Burlando, and J. D. Salas, Stochastic forecasting of precipitation and streamflow processes, in T. D. Potter and B. Colman (Eds.), *Handbook of Weather, Climate, and Water*, Chapter 34 John Wiley & Sons, Inc., New York, 2002.

Valencia, R. D., and J. C. Schaake, Jr., Disaggregation process in stochastic hydrology, *Water Resour. Res.*, *20*(1), 580–585, 1973.

Visser, H., and J. Molenaar, Trend estimation and regression analysis in climatological time series: An application of structural time series models and the Kalman filter, *J. Climate*, *8*, 969–979, 1995.

Vorosmarty, C. J., P. Green, J. Salisbury, and R. B. Lammers, Global water resources: Vulnerability from climate change and population growth, *Science*, *289*, 284–288, 2000.

Wallis, J. R., and P. E. O'Connell, Firm reservoir yield: How reliable are hydrological records, *Hydrol. Sci. Bull* (now *Hydrol. Sci. J.*), *18*, 347–365, 1973.

Xu, J. S., and H. V. Storch, Predicting the state of the Southern Oscillation using principal oscillation pattern analysis, *J. Climate*, *3*, 1316–1329, 1990.

Yevjevich, V., An objective approach to definitions and investigations of continental droughts, in *Hydrology Papers*, Vol. *23*, Colorado State University, Fort Collins, CO, 1967.

Yevjevich, V., Structural analysis of hydrologic time series, in *Hydrology Papers*, Vol. 56, Colorado State University, Fort Collins, CO, 1972a.

Yevjevich, V., *Stochastic Process in Hydrology*, Water Resources Publications, Littleton, CO, 1972b.

Zheng, X., and R. E. Basher, Structural time series models and trend detection in global and regional time series, *J. Climate*, *12*, 2347–2358, 1999.

CHAPTER 33

STOCHASTIC SIMULATION OF PRECIPITATION AND STREAMFLOW PROCESSES

JOSÉ D. SALAS, JORGE A. RAMÍREZ, PAOLO BURLANDO, AND ROGER A. PIELKE, Sr.

Stochastic simulations of hydroclimatic processes such as precipitation and streamflow have become standard tools for analyzing many water-related problems. Simulation signifies "mimicking" the behavior of the underlying process so that realistic representations of it can be made. For this purpose a number of empirical, mathematically/physically based, mathematically/stochastically based, analog/physically based, and physical/laboratory-scale based models and approaches have been proposed and developed in the literature. This chapter emphasizes simulation based on stochastic and probabilistic techniques. Also, the emphasis will be on precipitation and streamflow processes, although many of the methods and models included herein are equally applicable for other hydroclimatic processes as well such as evapotranspiration, soil moisture, surface and groundwater levels, and sea surface temperature.

Stochastic simulation enables one to obtain equally likely sequences of hydroclimatic processes that may occur in the future. They are useful for many water resources problems such as (a) estimating the design capacity of a reservoir system under uncertain streamflows, (b) evaluating the performance of a water resources system in meeting projected water demands under uncertain system's inputs, (c) estimating drought properties, such as drought length and magnitude based on simulated streamflows at key points in the water supply system under consideration, (d) deriving the distribution of the underlying output variable of a groundwater flow equation (e.g., the hydraulic head), given the distribution of the parameters (e.g., the hydraulic conductivity) and boundary conditions, (e) establish-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

ing the uncertainty in travel time and spread of pollutants in porous media as a function of the uncertainty in the parameters of the groundwater contamination transport model, and (f) analyzing the impacts of large-scale climate variability and global climate change on water supply availability and the ensuing planning and operation of water resources projects.

1 STOCHASTIC SIMULATION OF PRECIPITATION

Continuous-Time Precipitation

The theory of *point processes* has been applied for modeling continuous-time precipitation since Le Cam[60] suggested that a Poisson process could model the occurrence of rainfall showers. Let us assume that the number of storms $N(t)$ in a time interval $(0, t)$ arriving to a given point is Poisson distributed with parameter λt (λ = storm arrival rate.) Referring to Figure 1(*a*), n storms arrived in the interval $(0, t)$ at times $t_1, \ldots, t_n$. The number of storms in any time interval T is also Poisson distributed with parameter λT. Assume further that the rainfall amount R associated with a storm arrival is *white noise* (e.g., R may be gamma distributed) and that $N(t)$ and R are independent. Thus, rainfall amounts $r_1, \ldots, r_n$ correspond to storms occurring at times $t_1, \ldots, t_n$. Such a rainfall generating process has been called *Poisson white noise* (PWN).

The cumulative rainfall in the interval $(0, t)$, $Z(t) = \sum_{j=1}^{N(t)} R_j$ is a *compound Poisson process*. Also the cumulative rainfall over successive nonoverlapping time intervals T, i.e., the discrete-time rainfall process (refer to Fig. 1), is given by $Y_i = Z(iT) - Z(iT - T)$, $i = 1, 2, \ldots$. The basic statistical properties of Y_i assuming that $Z(t)$ is generated by a PWN model has been widely studied.[11,20] Its autocorrelation function $\rho_k(Y)$ is equal to zero for all lags greater than zero, which contradicts actual observations [e.g., $\hat{\rho}_1(Y) = 0.446$ for hourly precipitation at Denver Airport station for the month of June based on the 1948–1983 records.] Despite this shortcoming, the PWN model can be useful for predicting annual precipitation[20] and extreme precipitation events.[10] Instead of assuming that rainfall occurs instantaneously with zero duration one may consider rainfall with random duration D and intensity I, as shown in Figure 1*b*. This is called the *Poisson rectangular pulse* (PRP) model.[86] A common assumption is that D and I are independent and exponentially distributed. Figure 1*b* shows a PRP process with n storms in the interval $(0, t)$ occurring at times $t_1, \ldots, t_n$ with associated intensities and durations $(i_1, d_1), \ldots, (i_n, d_n)$. Then, storms may overlap and the aggregated process Y_i becomes autocorrelated. Although the PRP model is better conceptualized than the PWN, it is still limited when applied to rainfall data.[86] Thus, alternative models based on the concept of clusters have been suggested.

Neyman and Scott[69] in modeling the spatial distribution of galaxies originally suggested the concept of *clusters*. Le Cam[60], Kavvas and Delleur,[51] and others[30,82,86–88] applied this concept of space clustering to model continuous-time rainfall. The *Neyman–Scott cluster process* can be described as a two-level mechan-

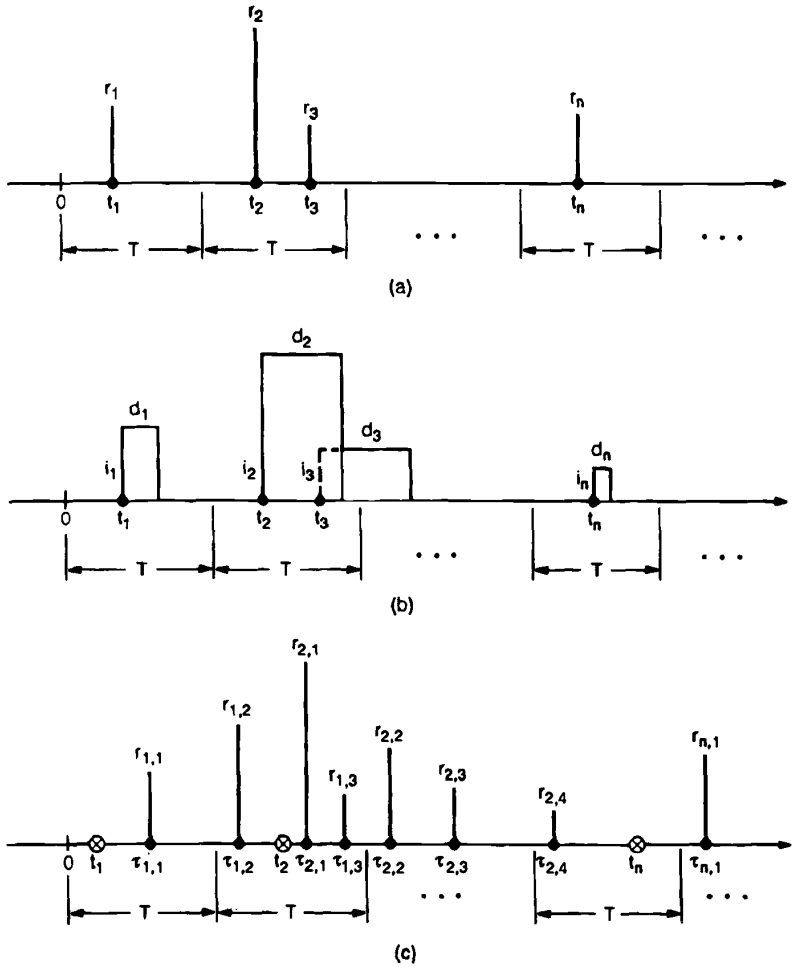


Figure 1 Schematic representations of (*a*) Poisson white noise, (*b*) Poisson rectangular pulse, and (*c*) Neyman–Scott white noise processes (after Salas[93]).

ism for generating rainfall. First, storm-generating mechanisms or simply storms arrive governed by a Poisson process with parameter λt. Figure 1*c* shows that n storms arrive at times $t_1, \ldots, t_n$ in the period $(0, t)$. Then, associated with each storm, there are a number of precipitation bursts that are Poisson or geometrically distributed with parameter v. Figure 1*c* shows three precipitation bursts associated with the storm that arrived at time t_1. In general m_j precipitation bursts are associated

with the storm that arrived at time t_j. In addition, the time of occurrence of bursts, τ, relative to the storm origin t_j may be assumed to be exponentially distributed with parameter β (e.g., in Fig. 1*c* the three bursts arising from the first storm are located at times $\tau_{1,1}$, $\tau_{1,2}$, and $\tau_{1,3}$ relative to t_1). Then, if the precipitation burst is described by an instantaneous random precipitation depth R, the resulting precipitation process is known as *Neyman–Scott white noise* (NSWN) while if the precipitation burst is a rectangular pulse the precipitation process is known as *Neyman–Scott rectangular pulse* (NSRP).

Estimation of parameters for Neyman–Scott (NS) models has been a major subject of research in the past two decades.[9,22,30,51,71] The usual estimation has been based on the method of moments, although other approaches have been suggested.[29,51,82] An apparent major estimation problem is that parameters estimated based on data for one level of aggregation, say hourly, may be significantly different from those estimated from data for another level of aggregation, say daily.[11,30,71,86] The problem seems to be that as data are aggregated, information is lost and corresponding second-order statistics do not have enough information to give reliable estimates of the parameters of the generating process (model), and, as a consequence, they become significantly biased with large variance. For example, extensive simulation studies were carried out by Cadavid et al.[11] based on the NSWN model with (known) population parameters: $\lambda = 0.102 \times 10^{-3}$/min, $\beta = 0.00221$/min, $\mu = 24.36$/in, and $1/\nu = 0.072$ (parameter of the geometric distribution for the cluster size). Hourly and daily series were used to estimate moments (mean, standard deviation, and lag-1 and lag-2 correlation coefficients) from which the parameters were estimated. The results are shown in Table 1. Clearly, despite that the generating mechanism is known (the NSWN), less reliable estimates of parameters are obtained when daily values are used. Estimation based on weighted moments of various time scales in a least-squares fashion is an alternative.[10,22] Also, physical considerations may be useful in setting up constraints in some of the parameters, initializing the estimates to be determined based on statistical considerations, and for comparing the fitted model parameters with some known physical properties.[17] Koepsell and Valdes[54] applied these concepts using the space–time cluster model suggested by Waymire et al.[116] for modeling rainfall in Texas and pointed out the difficulty in estimating the parameters even when using physical considerations.

Besides the class of Poisson processes and Neyman–Scott cluster processes, other types of temporal precipitation models have been suggested such as those based on Cox processes,[103] renewal processes,[7,31] and Barlett–Lewis processes.[40,87] Likewise, alternative space–time multidimensional precipitation models have been developed (e.g., Smith and Krajewski[104]). In addition, all precipitation models based on point and cluster processes proposed up to date are limited in some respects; e.g., they do not include the daily periodicity observed in actual convective rainfall processes.[49,71] Furthermore, Rodriguez-Iturbe et al.[89] and others raised the issue that nonlinear dynamics and chaos may be useful approaches for certain hydrometeorological processes such as rainfall. Finally, excellent reviews of the state of the art in the field have been made[32] and a number of studies pertaining to rainfall analysis, modeling, and predictability have been compiled in special issues of some

TABLE 1 Comparison between Population and Estimated Parameters of NSWN Model Based on Hourly and Daily Values

Parameter (units)	Population Value	Estimated from Hourly Data[a]	Estimated from Daily Data[a]
$\lambda \times 10^3$ (1/min)	0.102	0.103	0.091
$\beta \times 10^3$ (1/min)	2.210	2.300	1.630
μ (1/in)	24.360	23.990	7.010
$1/\nu$	0.072	0.072	0.247

[a]Estimates based on 12 series of size 36,456 for hourly and 12 series of size 1519 for daily. From Cadavid et al.[11]

journals (e.g., *J. Appl. Meteor.*, vol. 32, 1993; *J. Geoph. Res.*, vol. 104, no. D24, 1999).

Hourly, Daily, and Weekly Precipitation

We have seen in the previous section that the models and properties for cumulative precipitation over successive nonoverlapping time periods, i.e., discrete-time precipitation, can be derived from continuous-time precipitation models. However, one may formulate precipitation models directly at hourly, daily, and weekly time scales. In these cases, the theory of *Markov chains* has been widely used in the literature for simulating not only precipitation (in discrete time) but many other hydrologic processes such as streamflow, soil moisture, temperature, solar radiation, and water storage in reservoirs.[7,13,49,85,90]

Consider that $X(t)$ is a discrete valued process that started at time 0 and developed through time, i.e., $t = 0, 1, 2, \ldots$. Then $P[X(t) = x_t | X(0) = x_0,\ X(1) = x_1, \ldots, X(t-1) = x_{t-1}]$ is the probability that the process $X(t) = x_t$ given its entire history. If this probability simplifies to $P[X(t) = x_t \mid X(t-1) = x_{t-1}]$, the process is a *first-order Markov chain* or a *simple Markov chain*. Because $X(t)$ is a discrete valued process, we will use the notation $X(t) = j, j = 1, \ldots, r$ instead of $X(t) = x_t$, where j represents a *state* and r is the number of states; e.g., in modeling daily rainfall one may consider $r = 2$ with $j = 1$ for a dry day (no rain) and $j = 2$ for a wet day. A simple Markov chain is defined by its *transition probability matrix* $P(t)$, a square matrix with elements $p_{ij}(t) = P[X(t) = j | X(t-1) = i]$ for all i, j pairs. Furthermore, $q_j(t) = P[X(t) = j], j = 1, \ldots, r$, is the marginal probability distribution of the chain being at any state j at time t and $q_j(0)$ is the distribution of the initial states. Moreover, if $P(t)$ does not depend on time, the Markov chain is a *homogeneous* or *stationary chain* and, in this case, the notations P and p_{ij} are used. The estimation of some probabilities that are useful for simulation and forecasting of precipitation events are the n-step transition probability $p_{ij}^{(n)}$, the marginal distribution $q_j(t)$ given the distribution $q_j(0)$, and the steady-state probability vector q^*. These probabilities can be determined from well-known relations available in the literature.[39,118]

Estimation for a simple Markov chain amounts to estimating the elements p_{ij} of the transition probability matrix. Common estimation methods include the method of moments and maximum likelihood.[39] To test whether a simple Markov chain is an adequate model for the process under consideration, one can check some of the assumptions of the model and see whether some relevant properties of the precipitation process are reproduced (e.g., compare the probability $p_{ij}^{(n)}$ with that obtained from the observed data, $\hat{p}_{ij}^{(n)}$). Furthermore, the Akaike information criterion has been helpful in selecting the order of Markov chain models.[15,48]

Although in some cases simple Markov chains may be adequate for representing the variability of precipitation, often more complex models may be necessary. For instance, in modeling daily rainfall processes throughout the year, the parameters of the Markov chain may vary with time (e.g., for a two-state Markov chain, the transition probabilities p_{ij} may vary along the year and the estimates can be fitted with trigonometric series to smooth out sample variations[90]). Higher order Markov chains may be necessary in other cases. Chin[15] analyzed daily precipitation records of more than 100 stations across the continental United States and concluded that generally second- and third-order models were preferred for the winter months while the first-order model was better for the summer months. In addition, maximum likelihood for estimating Fourier series coefficients for alternating renewal processes and Markov chains for daily rainfall[90] and mixed models with periodic Markov chains for hourly rainfall (to account for the effect of daily periodicity) have been suggested.[49]

Monthly, Seasonal, and Annual Precipitation

Modeling of precipitation for long time scales such as monthly is generally simpler than for short time scales such as daily, especially because for long time scales the autocorrelation becomes smaller or negligible (except in cases of low frequency[21]). In such cases modeling precipitation at a given site amounts to finding the probability distribution for each month. Generally different distributions will be needed for each month. On the other hand, seasonal precipitation data in semiarid and arid regions may include zero values for some seasons, hence the precipitation is a mixed random variable. Let $X_{v,\tau}$ = precipitation for year v and season τ, and define $P_\tau(0) = P(X_{v,\tau} = 0)$, $\tau = 1, \ldots, \omega$ (ω = number of seasons per year). Then, $F_{X\tau}(x) = P_\tau(0) + [1 - P_\tau(0)]F_{X\tau|X\tau>0}(x)$ is the cumulative distribution function for season τ, in which $F_{X\tau}(x) = P_\tau(X \le x)$ and $F_{X\tau|X\tau>0}(x) = P(X \le x|X > 0)$. Thus, prediction of seasonal precipitation requires estimating $P_\tau(0)$ and $F_{X\tau|X\tau>0}(x)$. Several distributions such as the log-normal and log-Pearson have been used for fitting the empirical distribution of seasonal precipitation. For modeling precipitation at several sites, one must consider the intersite cross correlations and the marginal distribution (at each site). For continuous random precipitation, a common modeling approach has been to transform them into normal, then use a lag-0 multivariate model for modeling the transformed precipitation (an approach similar to modeling streamflow as in Section 2). Modeling of annual precipitation is similar to modeling seasonal precipitation, i.e., determining either the marginal distribution $F_X(x)$ or the

conditional distribution $F_{X*X>0}(x)$, depending on the particular case at hand. Likewise, modeling of annual precipitation at several sites is generally based on transforming the data into normal and using a multivariate normal model.

2 STOCHASTIC SIMULATION OF STREAMFLOW

If one can develop a stochastic model for streamflow in continuous time, then, in principle, the properties and the models for daily, monthly, and annual streamflow can be obtained. Some attempts have been made for developing models of streamflow processes in continuous time based on physical principles.[20] However, the models of aggregated flows that can be derived from such continuous-time models, become mathematically cumbersome and of limited applicability for operational hydrology.[53] Understanding the rules for upscaling the models and parameters has been a challenging subject for research. Generally most of the models that are available for streamflow simulation in continuous time and short time scales, such as hourly, are based on the transformation of precipitation into runoff by means of physical or conceptual principles. Thus the stochastic characteristics of the precipitation input and of the other relevant processes of the hydrologic cycle of the watershed are transferred into a stochastic streamflow output. Examples of models in this category are represented by SHETRAN[26] and PRMS.[59] SHETRAN simulates "continuous" streamflows along the river network by solving partial differential equations of the physical processes involved while PRMS is a semidistributed conceptual model that simulates hourly and daily streamflows. However, in this section we are mainly concerned with stochastic streamflow models that can be derived explicitly from the physically or conceptually based relations of the underlying hydrologic process of the watershed or directly from the streamflow data.

Continuous Time to Hourly and Daily Streamflow Simulation

The simulation of streamflow on a continuous time scale requires the formulation of a model structure that is capable of reproducing the streamflow fluctuations on a wide dynamical range. As already mentioned, the application of stochastic approaches to continuous time and short time scale streamflow modeling has been limited because of the complex nonlinear relations that characterize the precipitation-streamflow processes at those temporal scales. The early attempts to model hourly and daily streamflows were based on using autoregressive (AR) models after standardization and transformation. However, stochastic models essentially based on process persistence do not properly account the rising limb and recession characteristics that are typical of hourly and daily flow hydrographs. Also shot noise or Markov processes and transfer function models have been proposed for daily flow simulation with some limited success in reproducing the rising limb and recessions.[110]

Nevertheless, interesting work has been done with some success by using conceptual-stochastic models. For instance, Kelman[52] applied a conceptual representation

of a watershed considering the effects of direct runoff and surface and groundwater storages. Direct runoff is modeled by a PAR(1) model with an indicator function to produce intermittence and the other components are modeled using linear reservoirs. Kelman's model produced reasonable results for generating daily flows for the Powell River, Tennessee. Also following the approach suggested by Salas and Obeysekera[96] and Claps et al.,[16] Murrone et al.[68] proposed a conceptual-stochastic model for short time runoff. A three-level conceptual runoff component and a stochastic surface runoff model the daily response of the watershed. The base flow is modeled by three linear reservoirs that represent the contribution of deep aquifers with over-year response, aquifers with annual renewal, and subsurface runoff. The surface runoff is regarded as an uncorrelated point process. Modeling rainfall as an independent Poisson process, the above scheme leads to a multiple shot noise streamflow process. The model is effective in reproducing streamflow variability. In addition, intermittent daily streamflow processes have been modeled[25] by combining Kelman's conceptual approach with product models[94] and gamma AR models[28] and by using a three-state Markov chain describing the onset of streamflow and an exponential decay of streamflow recession.[1]

Weekly, Monthly, and Seasonal Streamflow

Single-Site Periodic Models. Stationary stochastic models can be applied for modeling weekly, monthly, and seasonal streamflows after *seasonal standardization*. This approach may be useful when the season-to-season correlations do not vary throughout the year. In general though, models with periodic correlation structure, such as periodic autoregressive (PAR) and periodic autoregressive and moving average (PARMA) are more applicable.[29,92] An example is the PARMA(1,1) model[97]

$$y_{v,\tau} = \mu_\tau + \phi_{1,\tau}(y_{v,\tau-1} - \mu_{\tau-1}) + \varepsilon_{v,\tau} - \theta_{1,\tau}\varepsilon_{v,\tau-1} \tag{1}$$

where μ_τ, $\phi_{1,\tau}$, $\theta_{1,\tau}$, and $\sigma_\tau(\varepsilon)$ are the model parameters. When the θ's are zeros, model (1) becomes the PARMA(1,0) or PAR(1) model. Low-order PARMA models such as PARMA(1,0) and PARMA(1,1) have been widely used for simulating monthly and weekly flows.[3,18,43,84,92,120]

PARMA models can be derived from physical/conceptual principles. Considering all hydrologic processes and parameters in the watershed varying along the year, it has been shown that seasonal streamflow falls within the family of PARMA models.[96] Alternatively, a constant parameter model with periodic independent residuals was suggested.[16] One of the desirable properties of stochastic models of seasonal streamflows is the preservation of seasonal and annual statistics. However, such dual preservation of statistics has been difficult to get with simple models such as the PAR(1) or PAR(2). For this reason in the 1970s hydrologists turned to the so-called *disaggregation* models (refer to Section 3). The major drawback of such simple PAR models to reproduce seasonal and annual statistics has been the lack of sufficient correlation structure. PARMA models having more flexible correlation

structure than PAR models offer the possibility of preserving seasonal and annual statistics. Some hydrologists have argued that PARMA models have too many parameters. Yet, one cannot hope for models such as the PAR(1) to do more than it can, i.e., to reproduce simply the lag-1 month-to-month correlations while failing to reproduce correlations for longer time lags and statistics at higher orders of aggregation. An alternative for reproducing both seasonal and annual statistics is the family of multiplicative models.

Box and Jenkins[5] first suggested multiplicative models. These models have the characteristic of linking the variable $y_{\nu,\tau}$ with $y_{\nu,\tau-1}$ and $y_{\nu-1,\tau}$. McKerchar and Delleur[66] used multiplicative models after differencing the logarithms of the original series for simulating and forecasting monthly streamflow series. Because such multiplicative models do not take into account periodic correlations, differencing was used in an attempt to decrease or eliminate such periodicity. However, they were not able to reproduce the seasonality in the covariance structure and could not establish confidence limits of forecasts with consideration of seasonality. This problem arises because the referred multiplicative model does not include periodic parameters. A model (with periodic parameters) that can overcome the limitations mentioned above is the multiplicative PARMA model.[95] For instance, the multiplicative PARMA(1,1) $\times$ $(1.1)_{\omega}$ model is written as

$$\begin{aligned} z_{\nu,\tau} = {} & \Phi_{1,\tau} z_{\nu-1,\tau} + \phi_{1,\tau} z_{\nu,\tau-1} - \Phi_{1,\tau}\phi_{1,\tau} z_{\nu-1,\tau-1} + \varepsilon_{\nu,\tau} \\ & - \Theta_{1,\tau}\varepsilon_{\nu-1,\tau} - \theta_{1,\tau}\varepsilon_{\nu,\tau-1} + \Theta_{1,\tau}\theta_{1,\tau}\varepsilon_{\nu-1,\tau-1} \end{aligned} \tag{2}$$

in which $z_{\nu,\tau} = y_{\nu,\tau} - \mu_{\tau}$ and $\Phi_{1,\tau}$, $\Theta_{1,\tau}$, $\phi_{1,\tau}$, $\theta_{1,\tau}$, and $\sigma_{\tau}(\varepsilon)$ are the model parameters. This model has been applied successfully for simulating the Nile River flows.

A limitation of the foregoing PARMA and multiplicative PARMA models for modeling hydrological time series is the requirement that the underlying series be transformed into normal. An alternative that does not have this requirement is the PGAR(1) model for modeling seasonal flows with periodic correlation structure and periodic gamma marginal distribution.[27] Consider that $y_{\nu,\tau}$ is a periodic correlated variable with a three-parameter gamma marginal distribution with location λ_{τ}, scale α_{τ}, and shape β_{τ} parameters varying with τ, and $\tau = 1, \ldots, \omega$ (T = number of seasons). Then, the new variable $z_{\nu,\tau} = y_{\nu,\tau} - \lambda_{\tau}$ is a two-parameter gamma that can be represented by $z_{\nu,\tau} = \phi_{\tau} z_{\nu,\tau-1} + (z_{\nu,\tau-1})^{\delta_{\tau}} w_{\nu,\tau}$ where ϕ_{τ} = periodic autoregressive coefficient, δ_{τ} = periodic autoregressive exponent, and $w_{\nu,\tau}$ = noise process. This model has a periodic correlation structure equivalent to that of the PAR(1) process. It has been applied to weekly streamflow series for several rivers in the United States.[27] Results obtained indicated that such PGAR model compares favorably with respect to the normal based models (such as the PAR model after logarithmic transformation) in reproducing the basic statistics usually considered for streamflow simulation. Furthermore, a nonparametric approach for streamflow simulation that is capable of reproducing closely historical distributions has been proposed.[100]

PARMA and PGAR models are less useful for modeling flows in ephemeral streams. In these streams the flows are intermittent, a characteristic that is not represented by the above mentioned models. Instead periodic product models such as $y_{v,\tau} = B_{v,\tau} z_{v,\tau}$ are more realistic,[94] where $B_{v,\tau}$ is a periodic correlated Bernoulli (1,0) process, $z_{v,\tau}$ may be either an uncorrelated or correlated periodic process with a given marginal distribution, and B and z are mutually uncorrelated. Properties and applications of these models for simulating intermittent monthly flows of some ephemeral streams have been reported in the literature.[14,94]

Multisite Periodic Models. In modeling seasonal streamflows at several sites, multivariate PAR and PARMA models are generally used.[6,42,92,97] For example, the multivariate PARMA(1,1) model is

$$Z_{v,\tau} = \Phi_\tau Z_{v,\tau-1} + \underline{\varepsilon}_{v,\tau} - \Theta_\tau \underline{\varepsilon}_{v,\tau-1} \tag{3}$$

in which $Z_{v,\tau} = Y_{v,\tau} - \underline{\mu}_\tau$; $\underline{\mu}_\tau$ is a column parameter vector with elements $\mu_\tau^{(1)}, \ldots, \mu_\tau^{(n)}$. Φ_τ and Θ_τ are $n \times n$ periodic parameter matrices, the noise term $\underline{\varepsilon}_{v,\tau}$ is a column vector normally distributed with $E(\underline{\varepsilon}_{v,\tau}) = \underline{0}$, $E(\underline{\varepsilon}_{v,\tau}\underline{\varepsilon}_{v,\tau}^T) = \Gamma_\tau$, and $E(\underline{\varepsilon}_{v,\tau}\underline{\varepsilon}_{v,\tau-k}^T) = 0$ for $k \neq 0$, and n = number of sites. In addition, it is assumed that $\underline{\varepsilon}_{v,\tau}$ is uncorrelated with $Z_{v,\tau-1}$. Parameter estimation of this model can be made by the method of moments, although the solution is not straightforward. Dropping the moving average term in (3), i.e., $\Theta_\tau = 0$ for all τ's, yields a simpler multivariate PARMA(1,0) or PAR(1) model. This simpler model has been widely used for generating seasonal hydrologic processes. Further simplifications of the foregoing models can be made to facilitate parameter estimation. Assuming that Φ_τ and Θ_τ of Eq. (3) are diagonal matrices, the multivariate PARMA(1,1) model can be decoupled into univariate models for each site. To maintain the cross correlation among sites $\underline{\varepsilon}_{v,\tau}$ is modeled as $\underline{\varepsilon}_{v,\tau} = B_\tau \underline{\xi}_{v,\tau}$ where $E(\underline{\xi}_{v,\tau}\underline{\xi}_{v,\tau}^T) = I$ and $E(\underline{\xi}_{v,\tau}\underline{\xi}_{v,\tau-k}^T) = 0$ for $k \neq 0$. This modeling scheme is a *contemporaneous* PARMA(1,1), or CPARMA(1,1), model. Useful references on this type of models are available in the literature.[42,84,92,97]

Annual Streamflows

Autoregressive (AR) and autoregressive and moving average (ARMA) models have been the most popular models for single site and multisite annual streamflow simulation. Specifically, low-order models have been widely applied for generating annual flow series.[29,42,61,62,72,92]

Single-Site Stationary Models. The AR(1) model is defined as $y_t = \mu + \phi(y_{t-1} - \mu) + \varepsilon_t$. Its autocorrelation function $\rho_k = \phi\rho_{k-1} = \phi^k$ decays exponentially as the time lag k increases. This model has been a prototype of *short memory* models because ρ_k goes to zero relatively fast and as a result $h \to \frac{1}{2}$ rather quickly in $E(R_n^{**}) \sim n^h$ (R_n^{**} = rescaled range of cumulative departures from the sample mean). A more versatile model than the AR(1) is the ARMA(1,1) given by[6,42,97]

$$y_t = \mu + \phi(y_{t-1} - \mu) + \varepsilon_t - \theta\varepsilon_{t-1} \tag{4}$$

Its autocorrelation function $\rho_k = (1 - \phi\theta)(\phi - \theta)(1 - 2\phi\theta + \theta^2)^{-1}\phi^{k-1}$ is more flexible than that of the AR(1) model because it depends on the two parameters ϕ and θ. The ARMA process can represent *long memory* dependence,[72,91] a property that is important for many rivers. AR and ARMA models assume that the underlying series is normally distributed, an assumption that is not always applicable for annual streamflow series. While one can circumvent this assumption by transforming the skewed series into an approximately normal series, a direct approach that does not require a transformation is a viable alternative. The *gamma autoregressive* (GAR) process $y_t = \lambda(1 - \phi) + \phi y_{t-1} + \eta_t$ offers such an alternative where y_t is gamma distributed with parameters λ, α, and β (the location, scale, and shape parameters, respectively), ϕ = autoregressive coefficient, and η_t = noise term. The GAR(1) model has the same autocorrelation function as that of an AR(1) model. Estimation procedures and applications of the GAR model for simulating annual streamflow series can be found in the literature.[28]

AR, ARMA, and GAR models are useful for modeling streamflow processes in perennial rivers, yet they are inadequate for *intermittent processes* such as streamflows in some ephemeral streams. Intermittent processes can be modeled as $y_t = B_t z_t$ where y_t = non-negative intermittent variable, B_t = dependent (1,0) *Bernoulli process*, z_t = positive valued continuous autocorrelated variable, for instance, an AR(1) process, and B_t and z_t are assumed to be mutually uncorrelated. Thus, the resulting product process y_t is intermittent and autoregressive. These models have been applied for modeling short-term rainfall and intermittent flow processes.[7,13,94]

Finally, other type of models, such as fractional Gaussian noise.[64] broken line,[6] shifting level,[93] and FARMA[42] have been proposed for representing certain special properties of annual streamflow time series. For example, the shifting level model has the capability of simulating time series with sudden changes, a property that has been observed in many hydroclimatic processes.

Multisite Stationary Models. Modeling of multiple time series is widely needed in hydrology. Consider the column vector Y_t with elements $y_t^{(1)}, \ldots, y_t^{(n)}$ in which n = the number of series (number of variables) under consideration. The multivariate AR(1) model is defined as[65]

$$Z_t = \Phi Z_{t-1} + \underline{\varepsilon}_t \tag{5}$$

in which $Z_t = Y_t - \underline{\mu}$, $\underline{\mu}$ is a column vector of means $\mu^{(1)}, \ldots, \mu^{(n)}$, $\underline{\varepsilon}_t$ is a column vector of normal noises $\varepsilon_t^{(1)}, \ldots, \varepsilon_t^{(n)}$, each with zero mean such that $E(\underline{\varepsilon}_t \underline{\varepsilon}_t^T) = \Gamma$ and $E(\underline{\varepsilon}_t \underline{\varepsilon}_{t-k}^T) = 0$ for $k \neq 0$, and Φ and Γ are $n \times n$ parameter matrices. In addition, it is assumed that $\underline{\varepsilon}_t$ is uncorrelated with Z_{t-1}. Model (5) is a prototype of short-memory models for multiple series and has been widely used in operational hydrology.[29,42,62,92] Likewise, the *multivariate* ARMA(1,1) model can be written as in Eq. (3) except that the parameters Φ and Θ do not depend on time.

Except for low-order multivariate AR models, using the full multivariate ARMA models often leads to complex parameter estimation.[73,92] Thus, model simplifications have been suggested. For instance, a *contemporaneous* ARMA (CARMA)

model results if Φ and Θ are diagonal matrices. This concept, which has been advocated by Salas et al.,[92] Stedinger et al.,[106] and Hipel and McLeod[42] can be extended to the general case. A contemporaneous relationship implies that only the dependence of concurrent values of the y's are considered important. Furthermore, the diagonalization of the parameter matrices allows "model decoupling" into component univariate models so that the model parameters do not have to be estimated jointly, and univariate modeling procedures can be employed. Thus, univariate ARMA(p, q) models are fitted at each site where each $\varepsilon_t^{(i)}$, $i = 1, \ldots, n$ is uncorrelated, but are contemporaneously correlated with a variance–covariance matrix Γ. Thus, the parameters, ϕ's and θ's in each model, can be estimated by using univariate estimation procedures and the ε's can be modeled by $\underline{\varepsilon}_t = B\underline{\xi}_t$ in which $\underline{\xi}$ is normal with $E(\underline{\xi}_t \underline{\xi}_t^T) = I$ and $E(\underline{\xi}_t \underline{\xi}_{t-k}^T) = 0$ for $k \neq 0$. Note that one does not have to consider the same univariate ARMA(p, q) model for each site.

3 TEMPORAL AND SPATIAL DISAGGREGATION MODELS

Disaggregation models, i.e., downscaling models in time and/or space, have been an important part of stochastic hydrology, not only because of our scientific interest in understanding and describing the characteristics of the spatial and temporal variability of hydrological processes, but also because of practical engineering applications. For example, many hydrologic design and operational problems require hourly precipitation data. Because hourly precipitation data are not as commonly available as daily data, a typical problem has been to *downscale* or *disaggregate* daily data into hourly data. Similarly, for simplifying the analysis and modeling of large-scale systems involving a large number of precipitation and streamflow stations, temporal and spatial disaggregation procedures are needed. This section briefly reviews some empirical and mathematical models and procedures for temporal and spatial disaggregation of precipitation and streamflow.

Disaggregation of Precipitation

Generally the disaggregation of station precipitation data defined at a given time interval into precipitation for smaller time intervals has been done empirically.[74] For instance, by using either tables or graphs, one can do disaggregation of 24-h (daily) precipitation into 6-h precipitation. More complete disaggregation schemes has been developed.[41,119]. Hershenhorn and Woolhiser[41] considered daily rainfall amounts and a model to obtain within-the-day magnitudes for the number of storms, amount, duration, and arrival time for each storm. They indicated that simulated rainfall sequences compared well with observed values. Although the foregoing models are innovative, they are not satisfactory, i.e., they are complex and require many transformations of the original data to obtain reasonable results. Another shortcoming is the lack of flexibility in the number of intervals considered.

Another formal disaggregating scheme for short-term rainfall was developed by Cadavid et al.[11] Disaggregation models were developed assuming PWN and NSWN

(refer to Section 1) as the underlying rainfall-generating mechanisms. Formulation of the disaggregation algorithm for the PWN model is based on the distribution of the number of arrivals N conditioned on the total precipitation Y in the time interval, the distribution of the white noise term given N and Y, and the distribution of the arrival times conditional on N. The algorithm performs well when using simulated PWN samples. The disaggregation scheme based on the NSWN model is more complex. It performs well on simulated and recorded samples provided that the model parameters used are similar to those controlling the process at the disaggregation scale. The main shortcoming is the incompatibility of parameter estimates at different aggregation levels as pointed out in Section 1. Recently, a rainfall disaggregation based on artificial neural networks has been suggested.[8]

Epstein and Ramírez[24] developed a multiscale, linear regression, statistical climate inversion scheme based on the disaggregation model of Valencia and Schaake[111] given by

$$Y = AX + B\varepsilon \tag{6}$$

where Y is a matrix of downscaled hydroclimatic values (e.g., precipitation), X is a matrix of upscaled hydroclimatic values, A and B are parameter matrices, and ε is a matrix of independent standard normal deviates. All terms in the above equation are functions of time, and the downscaling model is conditioned on time through the temporal evolution of the large-scale field, X. Parameter estimation, based on the method of moments, leads to the preservation of the first- and second-order moments at all levels of aggregation.

Disaggregation of Streamflow Data

The shortcoming of low-order PAR models when applied for simulating seasonal flows in reproducing the annual flow statistics led to the development of disaggregation models such as the Valencia–Schaake model (6). In this model, the modeling and simulation of seasonal flows is accomplished in two or more steps. First the annual flows are modeled so as to reproduce the desired annual statistics [e.g., based on the ARMA(1,1) model]; then synthetic annual flows are generated, which in turn are disaggregated into the seasonal flows by means of Eq. (6). While the variance–covariance properties of the seasonal flow data are preserved and the generated seasonal flows add up to the annual flows, model (6) does not preserve the covariances of the first season of a year and any preceding season. To circumvent this shortcoming, Eq. (6) has been modified as $Y = AX + B\underline{\varepsilon} + CZ$, where C is an additional parameter matrix and Z is a vector of seasonal values from the previous year (usually only the last season of the previous year) for each site.[6] Further refinements and corrections assuming an annual model that reproduces S_{XX} and S_{XZ} has been suggested[57] as well as a scheme that does not depend on the annual model's structure yet reproduces the moments S_{YY}, S_{YX}, and S_{XX}.[105]

The foregoing disaggregation models have too many parameters, a problem that may be significant especially when the number of sites is large and the available

historical sample size is small. Lane[56] sets to zero some of the parameters in the above disaggregation model so that

$$Y_\tau = A_\tau X + B_\tau \underline{\varepsilon} + C_\tau Y_{\tau-1} \qquad \tau = 1, \ldots, \omega \tag{7}$$

is a model with fewer parameters. Parameter estimation and appropriate adjustments so that the seasonal values add exactly to the annual values at each site can be found in the literature.[56,97]

The estimation problem can be simplified if the disaggregation is done in steps (stages or cascades) so that the size of the matrices involved and consequently the number of parameters decrease.[6] For instance, annual flows can be disaggregated into monthly flows directly in one step (this is the usual approach), or they can be disaggregated in two or more steps, e.g., into quarterly flows in a first step; then each quarterly flow is further disaggregated into monthly flows in a second step. However, even in the latter approach, considerable size of the matrices will result when the number of seasons and the number of sites are large. Santos and Salas[98] proposed a *stepwise disaggregation scheme* in such a way that at each step the disaggregation is always made into two parts or two seasons. This scheme leads to a maximum parameter matrix size of 2×2 for single-site disaggregation and $2n \times 2n$ for multisite. Disaggregation models that reproduce seasonal statistics and the covariance of seasonal flows with annual flows assuming log-normal seasonal and annual flows have been also suggested.[36,107] Also temporal disaggregation based on nonparametric procedures has been proposed.

Although disaggregation has been a major development and a practical tool in stochastic hydrology, still the question of why certain periodic models fail to reproduce annual statistics remained. Thus, more complex models such as PARMA models were suggested and developed in the 1970s and the early 1980s. Their capabilities for reproducing statistical properties beyond the seasons have been explored by Obeysekera and Salas[70] and Bartolini and Salas.[3]

4 TEMPORAL AND SPATIAL AGGREGATION MODELS

As in disaggregation, the *aggregation* (*upscaling*) modeling approach deals with streamflow processes at two or more levels of aggregation or time scales. However, the two concepts are quite different. In disaggregation, the modeling and generation procedure is backward in the sense that one models and generates annual flows first, then monthly, weekly, and daily flows are obtained in successive disaggregation steps. On the other hand, in temporal aggregation, the procedure is forward, i.e., one models and generates daily flows first and then successively weekly, monthly, and annual flows are modeled and generated. The basic premise of the aggregation approach is that the stochastic characteristics at the continuous time scale dictate those at any level of aggregation or time scale. The relationship between statistics at various time scales has been explored in the literature.[50]

The aggregation modeling approach for streamflow processes was developed by Vecchia et al.[112] Assuming that monthly flows follow a PAR(1) or PARMA(1,1) process, it was shown that the resulting model for the annual flows is the stationary ARMA(1,1). The foregoing concepts and results brought into light the structural linkage and compatibility between streamflow models (and their parameters) of various time scales. Streamflow data of the Niger River at Kaulikoro, Africa, were used to illustrate some of the aggregation concepts, especially in relation to reproducing the annual correlation structure when the seasonal flows were modeled by the PAR(1) and PARMA(1,1) models.[70] It was shown that in comparing the parameters and the correlograms of the ARMA(1,1) models of annual flows derived from the models of seasonal flows, the results obtained from the PARMA(1,1) model were significantly better than those obtained from the PAR(1). In addition, the results obtained vary depending on the number of seasons considered in the year (e.g., monthly, quarterly), and better results were obtained, as the number of seasons in the year became smaller.

Bartolini and Salas[3] extended the concept of aggregation to include aggregation not only from seasons to a year but from weeks to months, months to seasons, and seasons to years. For instance, the aggregation of a PARMA(2,1) monthly flows leads to a PARMA(2,2) bimonthly flows; in turn the aggregation of such PARMA(2,2) bimonthly flows gives also a PARMA(2,2) for the quarterly flows. Furthermore, if such quarterly flows are aggregated into annual flows, then the model is the stationary ARMA(2,2). The partial aggregation concepts have been applied to the Niger River seasonal and annual flows, and the results showed the superiority of the PARMA(2,1) and PARMA(2,2) models relative to the other models tested in reproducing the variance–covariance properties of the annual flows. The application of the aggregation concepts for modeling the seasonal and annual flows of the Niger River at various time scales suggest the need for using PARMA models for streamflow modeling and simulation if one would like to reproduce seasonal and annual first- and second-order statistics. The traditional models such as the PAR(1) simply are deficient for modeling flows such as those of the Niger. The usual approach to model seasonal and annual flows has been to use different models for different time scales, disregarding the compatibility among models at various time scales. The aggregation concepts and results discussed in this section point out that such traditional approaches and models for streamflow simulation must be avoided.

Similar reasoning as in temporal aggregation applies for spatial aggregation of streamflow processes. For instance, one can assume a stream network composed of a number of first-, second-, and third-order streams. Consider first modeling the flows at the junction of two first-order streams. Naturally, the model at a site immediately downstream from the junction must be derived from the bivariate model defined at two upstream sites, one in each of the tributaries. In turn, the model of the flows of the second-order stream as it joins the flows of another, say, second-order stream must define the model of the flows immediately downstream from the junction, and so on as the streamflow travels down the stream network. Thus, streamflow models must be compatible in both temporal and spatial scales.

5 SCALING ISSUES AND DOWNSCALING

Understanding, describing, and modeling local, regional, and global climate and its nonlinear interactions with hydrological, biophysical, and biogeochemical processes are currently some of the most challenging problems in the geosciences. It is not only the high spatial and temporal variability of the governing processes and boundary conditions but the wide range of scales over which this variability occurs. Distributed hydrologic models require high-resolution input data. Among all hydrologic variables, precipitation is of paramount importance in the water and energy budgets at the land surface–atmosphere interface, and its accurate representation in hydrologic and atmospheric models is critical. Precipitation is the result of intricately interrelated atmospheric and land surface processes. It has extreme variability over time scales from seconds to years and spatial scales from less than meters to hundreds of kilometers. The sensitivity of hydrologic behavior to the space–time variability of rainfall is the result of nonlinear interactions between precipitation and land surface characteristics controlling the transformation of rainfall into soil water and runoff. Modeling of precipitation requires understanding of the statistical structure of space–time precipitation and understanding of the physical processes governing the evolution of precipitation at a range of space–time scales. Because of scale differences, rainfall downscaling is required for coupling global (or regional) atmospheric models and hydrologic models. In general, downscaling schemes can be grouped in two broad categories, *dynamical* and *statistical* downscaling schemes.

In *dynamic* schemes, climate and land-use change scenarios at regional and local scales are developed using regional and local atmospheric models, for example, the Regional Atmospheric Modeling System (RAMS) of Colorado State University. These models are driven by boundary conditions derived from observations and from the output of global atmospheric models. In this way, the atmospheric model acts as a physically based dynamic interpolator (i.e., physically based downscaling). RAMS has been coupled to a land surface scheme (LEAF-2), to a hydrologic model (TOPMODEL), and to a regional ecosystem model (CENTURY). Thus, dynamical schemes encode multiple, nonlinear and complex local and regional interactions and feedbacks, explicitly.[79,115] However, trying to resolve processes at ever decreasing scales using physically based models rapidly leads to computational inefficiencies and is limited by poor understanding of physical process behavior at small scales. Other atmospheric models such as the NCAR models[33] and the Penn State/NCAR mesoscale model, Version 5,[19] have been used for dynamic downscaling.

In *statistical downscaling*, subgrid temporal and spatial scale details of climatic variability, in particular precipitation, are obtained so that the statistical characteristics of the spatial and temporal variability of hydroclimatic fields is preserved as a function of scale. *Statistical* techniques are commonly based on linear or nonlinear regression, methods from nonlinear dynamics, artificial neural networks, Markov processes, multiplicative random cascade models, etc. One of the limitations of regression approaches is that they are applicable only if a strong relationship between a large-scale parameter and regional and local climate has been identified (often this will not be the case), and they are valid only within the spatial and

temporal range of the observations. Although statistical downscaling is computationally efficient, it cannot include the subgrid scale physical feedbacks referred to above explicitly, and it is difficult to couple atmospheric processes with regional ecological and hydrological processes. On the other hand, statistical downscaling based on multiplicative random cascade models can reproduce the scaling features (i.e., scale invariance), the clustering, and intermittency that are characteristic of precipitation fields in space and time with relatively modest computational burden.

Until recently, most of the downscaling methods proposed in the literature only dealt with the spatial variability of the precipitation field while the temporal evolution of the fields is usually described independently of the spatial downscaling. The only temporal correlation structure accounted for is that resulting from the dynamics of the atmospheric model producing the precipitation field at the larger spatial scale or that encoded in the temporal evolution of the observations. Thus, in general, these schemes do not fully and properly account for the temporal correlation structure (i.e., persistence) of the precipitation fields at subgrid scales.

Dynamical Downscaling

Dynamical downscaling can be considered with respect to four basic types of models: one type is strongly dependent on larger-scale numerical weather prediction lateral boundary conditions, bottom boundary conditions, and on initial conditions. A second type has forgotten the initial conditions but is dependent on the observed lateral and bottom boundary conditions. A third type is where a large-scale model is run that is only forced with surface boundary conditions, and the output used to downscale with a regional model. A fourth type is when a true global climate model (with coupled ocean, atmosphere, continental sea ice, landscape processes, etc.) is used to provide lateral boundary conditions to a regional model. This is the Intergovernmental Panel on Climate Change (IPCC) type of downscaling except only a limited set of Earth system forcings (e.g., the radiative effect of CO_2, solar insolation) is included in the IPCC approach. To summarize with examples (IC = initial conditions; LBC = lateral boundary conditions, and BBC = bottom boundary conditions; with the recognition that BBC includes bottom interfacial fluxes): type 1 ETA[4] uses observed IC, LBC, and BBC; type 2 PIRCS[34,35] uses observed LBC and BBC; type 3 ClimRAMS forced by CCM3 integrated with observed SSTS[102] uses observed BBC; and type 4 Earth system global model downscaled using a regional model.[45] Observational constraints on the solution become less as we move from type 1 to 4. Thus forecast skill will diminish from type 1 to 4.

With respect to current generation models, such as atmospheric–ocean global circulation models (AOGCMs), neither AOGCMs nor the regional ones (type 4 models) include all of the significant human effects on the climate system. The combined effects of land-use change, the biogeochemical effect on the atmosphere, e.g., due to increased CO_2, and, e.g., the microphysical effect of pollution aerosols have not yet been included in these models. Thus the existing model runs should only be interpreted as sensitivity experiments, not forecasts, projections, or even scenarios.[80]

In addition, with respect to dynamic downscaling, as currently applied, there is not a feedback upscale to the AOGCM from the regional model, even if all of the significant large-scale (GCM scale) human-caused disturbances were included. The AOGCM also has a spatial resolution that is inadequate to properly define the lateral boundary conditions of the regional model. Anthes and Warner[2] show that the lateral boundary conditions are the dominant forcing of regional atmospheric models as associated with propagating features in the polar westerlies. With numerical weather prediction (type 1 and 2 models), the observations used in the analysis to initialize a model retain a component of realism even when degraded to the coarser model resolution of a global model. This realism persists for a period of up to a week or so, when used as lateral boundary conditions for a regional numerical weather prediction model. This is not true with the AOGCMs where data do not exist to influence the predictions. A regional model cannot reinsert model skill, when it is so dependent on lateral boundary conditions, no matter how good the regional model.

Statistical Downscaling

The output of mesoscale atmospheric models such as RAMS or the observations data such as those from the NEXRAD (next generation radar) network are usually at grid sizes that are larger [e.g., $O(10^3$ to $10^4)$ m] than those associated with distributed hydrologic models (e.g., in the order of 10^0 m). The land surface system responds to excitations from the atmosphere, e.g., precipitation, and feeds back moisture into the atmosphere, e.g., through evapotranspiration and latent heat flux. The excitations and responses are spatially heterogeneous over a broad range of scales, including subgrid scales [e.g., $< O(10^4)$m]. This subgrid spatial variability has a significant impact on the magnitude and distribution of upscale and downscale land surface fluxes whose interaction is nonlinear. Accounting for this space–time heterogeneity is important for hydrologic modeling and for describing land surface–atmosphere interactions.[23,78,83] In addition, the statistical downscaling, besides requiring that the AOGCMs are accurate predictions of the future, also require that the statistical equations used for downscaling remain invariant under changed regional atmospheric and land surface conditions. There is no way to test this hypothesis. In fact, it is unlikely to be valid since the regional climate is not passive to larger-scale climate conditions but is expected to change over time and feedback to the larger scales. More details of this concern regarding downscaling have been reported.[81]

Regression Schemes. The relationships between large-scale and local-scale climatic fields can be established by regression-based schemes. The most direct way of downscaling is by direct interpolation. This method is easy to apply and effective for smoothly varying fields such as sea level pressure or temperature but not appropriate for nonsmooth intermittent fields such as precipitation. Some examples on regression schemes are (1) a method (based on principal component analysis, canonical correlation, and regression analysis) called climatological projection by

model statistics to relate general circulation model (GCM) grid point free atmosphere statistics to important surface observations;[47] (2) use of interannual variations in climate to derive, through conventional regression analysis, statistical relationships between large-scale climate variations and local values of temperature and precipitation;[117] (3) a multiscale, linear regression, statistical climate inversion scheme that preserves all first- and second-order moments across scales (spatial downscaling of precipitation and temperature are applied in the context of impact assessment studies associated with global climate variability[24]); and (4) methods for conditional stochastic generation of rain fields used for disaggregation when constrained to large-scale values that are given by some outer sources.[58, 67]

Scale Invariance Scheme. Self-similarity or scale invariance is a kind of symmetry observed in nature. *Simple scaling* is a scaling in the probability distributions. Letting $R(\ )$ be the rainfall intensity field, this property is expressed as $R(\lambda A) \overset{\text{dist}}{=} \lambda^{-\theta} R(A)$, which indicates that the probability density function of the rescaled variable $R(\lambda A)$ is equal to that of the original variable $R(A)$ except for a factor that is a function of the length scale ratio λ and the scaling exponent θ. Simple scaling translates into two properties: (1) a log-log linearity between statistical moments of order n and length scale λ, i.e., $E[R_\lambda^n] = \lambda^{\theta(n)} E[R_1^n]$ and (2) a linear dependence on n of the slope $\theta(n)$ of that log-log linear relationship, i.e., $\theta(n) = n\theta$. Simple scaling is associated with additive (linear) processes, and unique scaling exponents θ are related to unique fractal dimensions. For example, Figures 2 and 3 are the scaling plots for the NEXRAD rainfall scan of July 4, 1997, for the central United States that is produced at grid scales of 2 km×2 km.[46] They show that the precipitation data for this date and region follow a simple scaling. It is often found that property (1) holds but the slope function $\theta(n)$ is nonlinear, a structure called *multiscaling* or *multifractal*.[37] Multifractal scaling behavior (i.e., scale-invariant

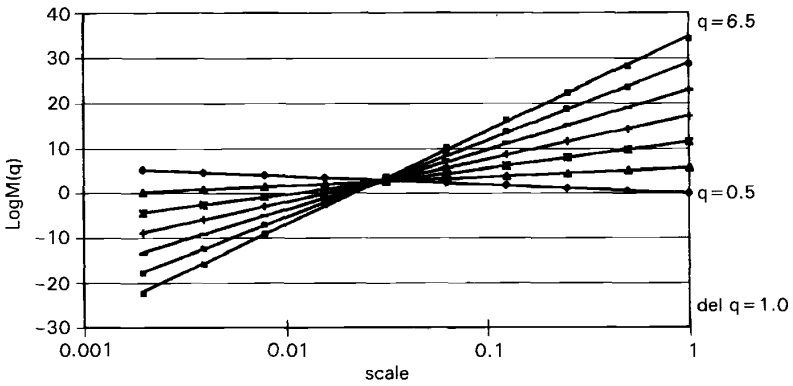


Figure 2 Marginal moments of precipitation over the central United States on July 4, 1997, from NEXRAD scan at 2 km×2 km grid size (from Kang and Ramirez[46]).

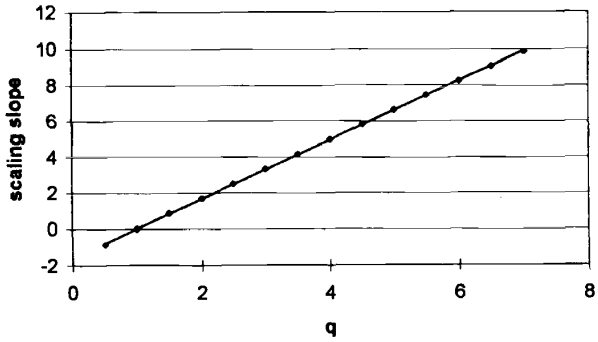


Figure 3 Slope of marginal moments log-log scaling function of NEXRAD scan of precipitation over the central United States on July 4, 1997, at 2 km×2 km grid size (from Kang and Ramirez[46]).

behavior) of the scaling exponents has been found in the spatial distribution of rainfall[37,75,109] and in the temporal distribution of rainfall.[12] Also both the scale invariance and intermittency of precipitation may be exploited to develop parsimonious stochastic models of rain.[63]

Multiplicative random cascades have been used to generate fractal fields that emulate the spectrum of scaling exponents of observed rainfall. Cascade generators are chosen according to the scaling spectra they produce. Notably, models such as the *universal multifractals*,[109] the β-model[38] or the log-Poisson model[101] were proposed. For illustration, we will discuss below random cascade models.[38,46,75]

Discrete random cascades distribute mass on successive regular subdivisions of a d-dimensional cube. A schematic of this process is shown in Figure 4. The initial cube, with length scale L_0, is subdivided at each level into b equal parts, where $b \geq 2^d$ is the branching number. The ith subcube after n levels of subdivision is denoted Δ_n^i (there are $i = 1, \ldots, b^n$ subcubes at level n). The length scale of the subcube Δ_n^i is denoted L_n, and the dimensionless spatial scale is defined as $\lambda_n = L_n/L_0 = b^{-n/d}$. The distribution of mass through different levels on the cubes occurs as follows. First the initial cube (at level $n = 0$) is assigned a nonrandom density R_0, i.e., an initial mass $R_0 L_0^d$. The subcubes Δ_1^i, $i = 1, \ldots, b$ after the first subdivision (at level $n = 1$) are assigned the density $R_0 W_1(i)$, i.e., mass $R_0 L_1^d W_1(i)$, where W are independent and identically distributed (iid) random variables—the *cascade generator*. This multiplicative process continues through all n levels of the cascade, so that the mass in subcube Δ_n^i is

$$\mu_n(\Delta_n^i) = R_0 L_n^d \prod_{j=1}^{n} W_j(i) = R_0 L_0^d b^n \prod_{j=1}^{n} W_j(i) \qquad \text{for } i = 1, 2, \ldots, b^n$$

where $E[W] = 1$; thus mass is on the average conserved at all levels in the random cascade.

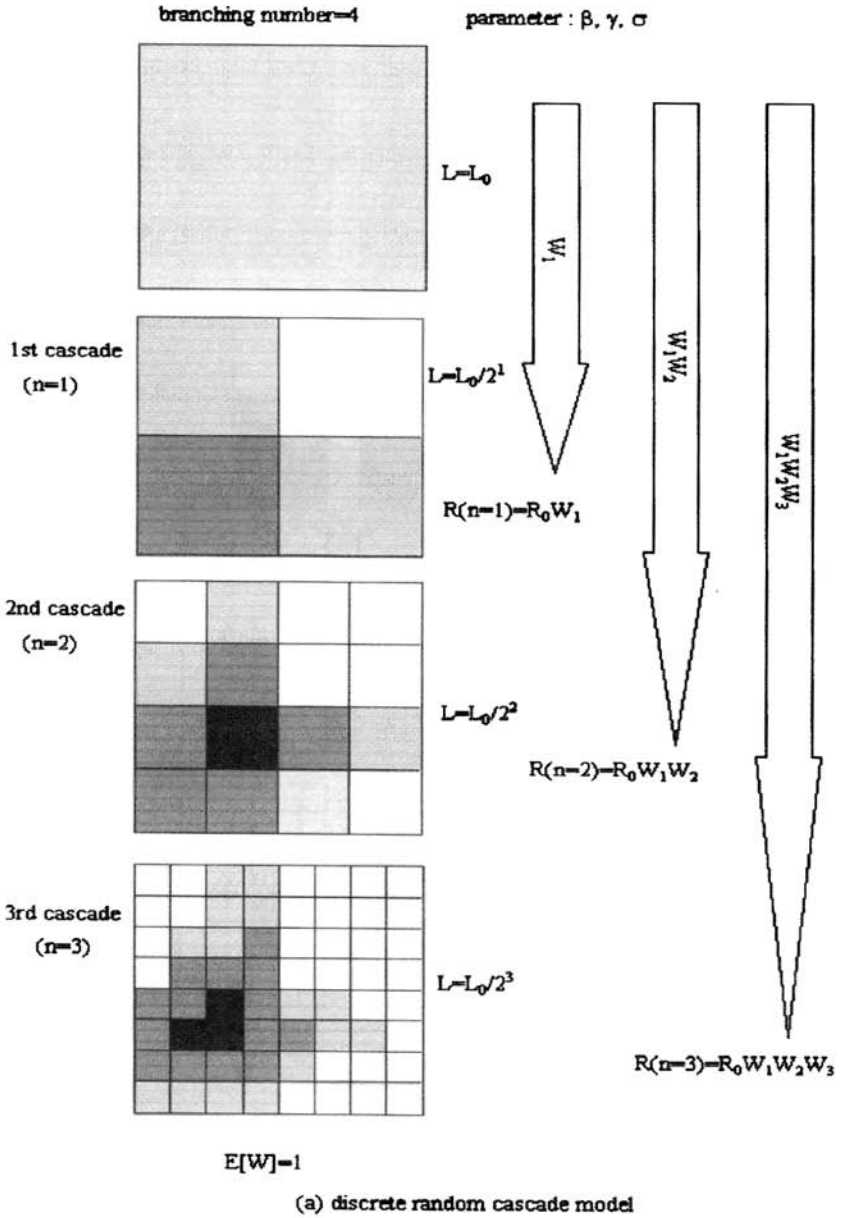


Figure 4 Schematic diagram for a two-dimensional discrete random cascade models: (*a*) Discrete random cascade model (from Kang and Ramirez[46]).

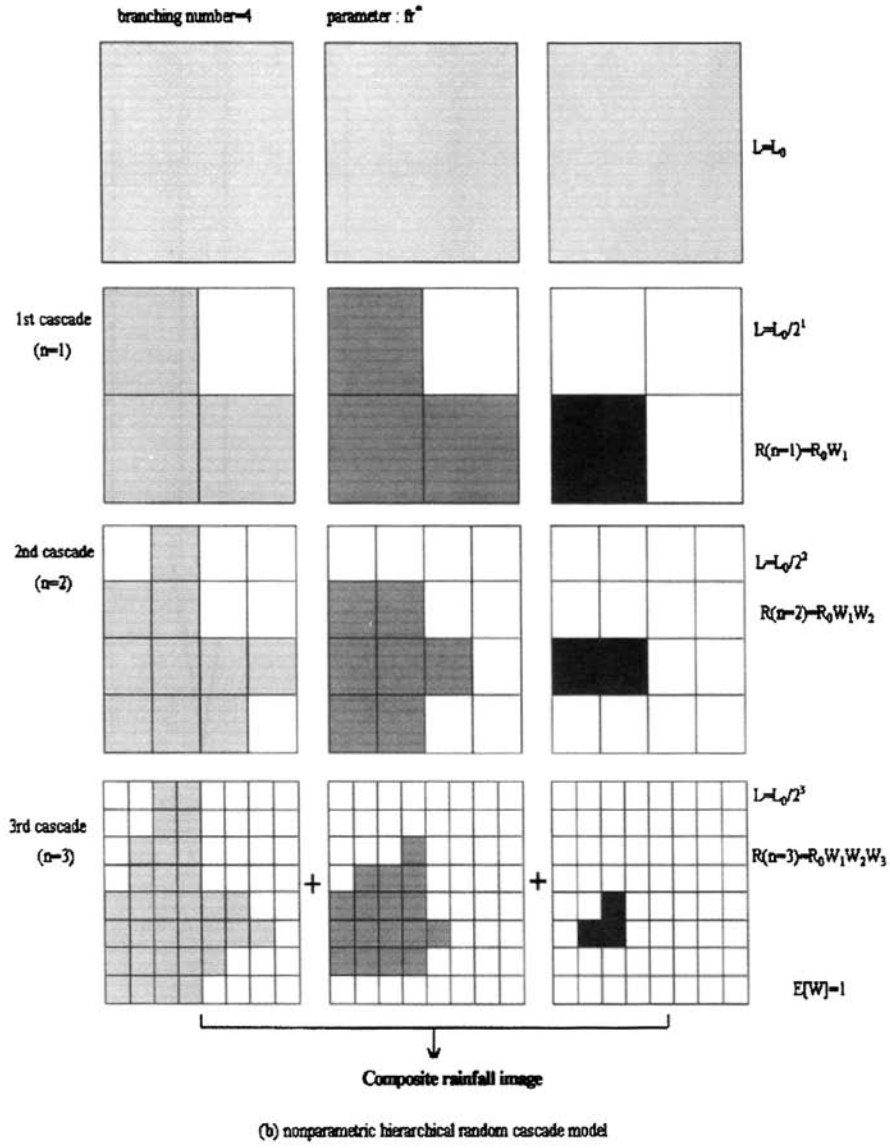


Figure 4 Schematic diagram for a two-dimensional discrete random cascade models: (*b*) nonparametric random cascade model (from Kang and Ramirez[46]).

The cascade limit mass μ_∞ is obtained as $n \to \infty$, and it is considered degenerate if the total mass is zero with probability 1. Nondegeneracy depends on the distribution of W, and it requires that the condition $E[W] = 1$ be satisfied. The limit mass in a subcube $\mu_\infty(\Delta_n^i)$ satisfies a recursion relation:[75] $\mu_\infty(\Delta_n^i) = \mu_n(\Delta_n^i)Z_\infty(i)$, for $i = 1, \ldots, b^n$ where Z_∞ are iid random variables, distributed as $Z_\infty = \mu_\infty(\Delta_0)/$

$\mu_0(\Delta_0) = \mu_\infty(\Delta_0)/R_0L_0^d$ for all i, n. The cascade limit mass $\mu_\infty(\Delta_n^i)$ is given by the product of a large-scale low-frequency component $\mu_n(\Delta_n^i)$, and a subgrid subgrid (i.e., subcube) scale high-frequency component $Z_\infty(i)$. The latter term represents subgrid-scale variability at each level of cascade development.

Random cascades exhibit *moment scaling* behavior from which properties of the cascade generator W can be estimated. Sample spatial moments are defined as $M_n(q) = \sum_{i=1}^{b^n} \mu_\infty^q(\Delta_n^i)$, where q = moment order (for $q = 0$, only nonzero limit masses are included in the sum). For large n, the sample moments should converge to the ensemble moments, but since they diverge to infinity or converge to zero as $n \to \infty$, the rate of convergence/divergence of the moments with scale is considered instead. In a random cascade, the ensemble moments are shown to be a log-log linear function of the scale λ_n. The slope of this scaling relationship is known as the Mandelbrot–Kahane–Peyriere (MKP) function: $\chi_b(q) = 1 - q + \log_b E[W^q]$. The MKP function contains important information about the distribution of the cascade generator W and thus characterizes the scaling properties of rainfall. Similarly, the slope of the sample moment scaling relationship can be defined as $\tau(q) = \lim_{\lambda_n \to 0}[\log M_n(q)/ - \log \lambda_n]$. For large n (as $\lambda_n \to 0$) and for a specific range of q, slopes of the moment scaling relationships for sample and ensemble moments converge, i.e., $\tau(q) = d\chi_b(q)$. In data analysis, the scaling of the sample moments is used to estimate the $\tau(q)$ function and the distribution of the cascade generator, from which parameters of the cascade model can be inferred.

For intermittent temporal and spatial rainfall data, it is desirable that $P(W = 0)$ be positive. For this purpose an intermittency model for the cascade generator W is written as $W = BY$, where B is an intermittency generator of the so-called β-model and Y is a strictly positive random variable. The β model divides the domain into rainy and nonrainy fractions based on the following probabilities: $P(B = 0) = 1 - b^{-\beta}$ and $P(B = b^\beta) = b^{-\beta}$, where β is a parameter and $E[B] = 1$. The β model does not allow for variability in the positive part of the large-scale component of the limit mass $\mu_n(\Delta_n^i)$ (at every level n it assumes the nonrandom value $R_0L_ndb^{\beta n}$). Variability in the positive part of the limit mass is obtained from the second element in the composite generator Y. The distribution of Y is arbitrary, but it has to be positive and $E[Y] = 1$. For rainfall modeling, good results have been obtained with Y log-normal.[75] Consider $Y = b^{\gamma+\sigma X}$, where X is a normal $N(0, 1)$ random variable. The condition $E[Y] = 1$ gives $Y = b^{-\sigma^2 \ln b/2+\sigma X}$, where σ^2 = variance of Y. Then W is distributed as $P(W = 0) = 1 - b^{-\beta}$ and $P(W = b^\beta Y = b^{\beta-\sigma^2 \ln b/2+\sigma X}) = b^{-\beta}$ with parameters β and σ^2.

The MKP function of W is $\chi_b(q) = (\beta - 1)(q - 1) + (\frac{1}{2})\sigma^2 \ln b(q^2 - q)$. MKP functions must be convex and $\chi_b(1) = 0$.[75] Also for the cascade to be nondegenerate it is required that $\chi_b^{(1)} < 0$. For large σ, this requirement may not be met, leading to degenerate cascades. The range of moments q in nondegenerate cascades is given by $[0, q_c/2)$ when $q_c \leq 2$, and at least in the closed interval $[0,1]$ when $q_c > 2$. The latter condition implies that $\beta < 1 - \sigma^2 \ln b$ in nondegenerate cascades.[75] Figure 5 shows the relationship between the parameters of the random cascade model and the large-scale forcing, here described in terms of the large-scale average rainfall intensity, for the NEXRAD data.[46]

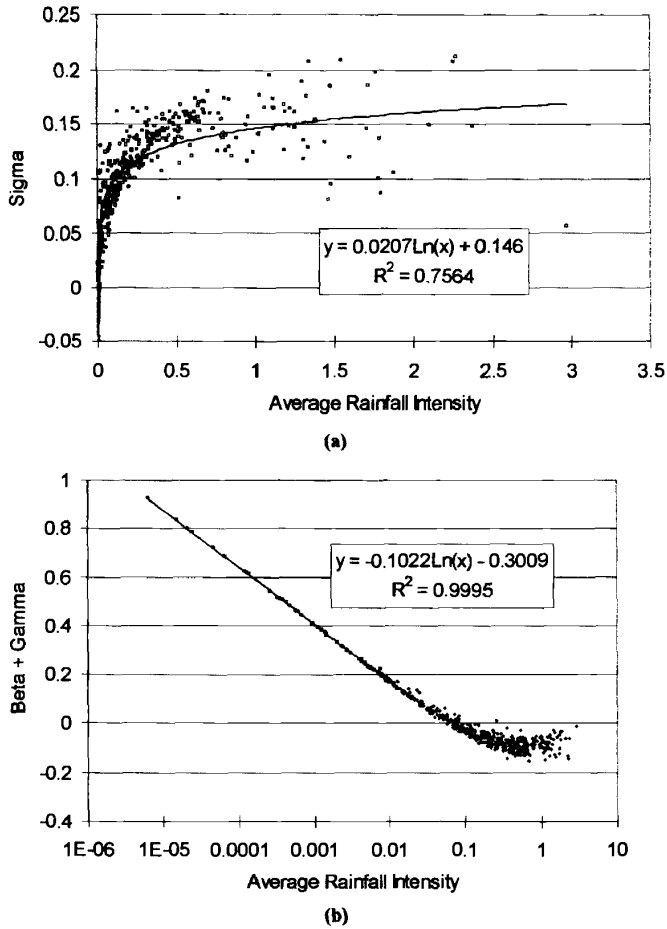


Figure 5 (*a*) Sigma vs. average rainfall intensity and (*b*) Beta + Gamma vs. average rainfall intensity (from Kang and Ramirez[46]). See ftp site for color image.

Because of the nondegeneracy condition on β, the β model is limited in the range of fractional rainy area values that it can describe. To deal with this limitation and to improve the preservation of the clustering structure, a simple pragmatic modification of the β model has been proposed, and the resulting scheme is referred to as a *nonparametric hierarchical* scheme. This procedure performs spatial downscaling by hierarchically applying a sequence of random cascade models to a set of precipitation fields defined in terms of intensity classes of the observations. Figure 6 shows a comparison of observed and simulated precipitation fields, using the β-log-normal random cascade model and the nonparametric hierarchical scheme. Figures 7

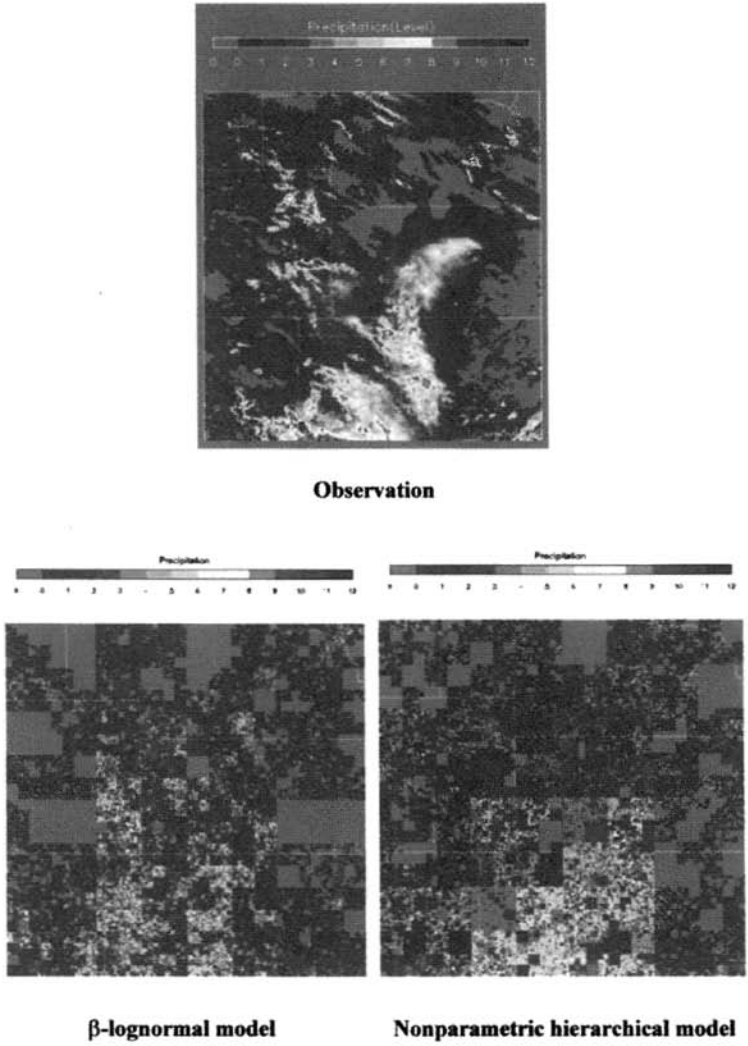


Figure 6 (see color insert) Comparison of observed and downscaled rainfall fields (July 6, 1997) (from Kang and Ramirez[46]). See ftp site for color image.

to 9 show a comparison of the ability of the respective models to reproduce important characteristics of the precipitation field. Table 2 presents a comparison of some summary statistics.

Further Remarks

The parameters of the scaling characteristics of the precipitation field must be related to some physically meaningful and measurable variable characterizing the large-

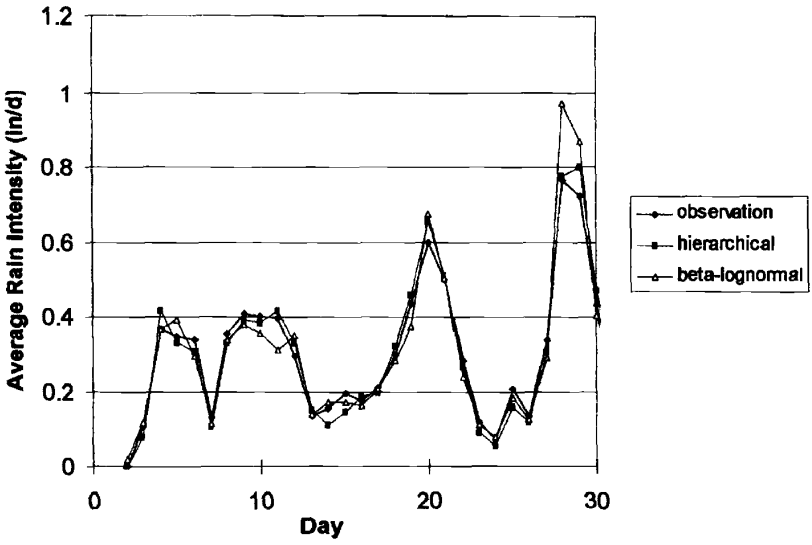


Figure 7 Time series of average rainfall intensity (NEXRAD, July 1997) (from Kang and Ramirez[46]). See ftp site for color image.

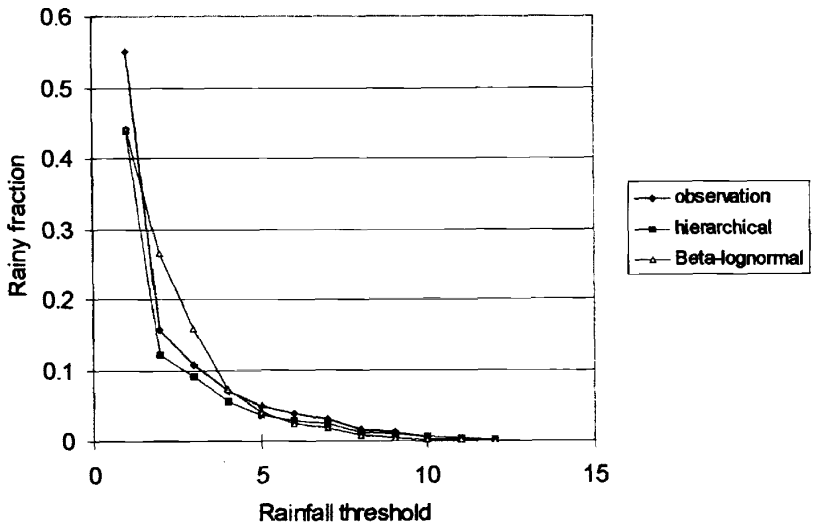


Figure 8 Rainy fraction as a function of rainfall threshold (NEXRAD, July 6, 1997) (from Kand and Ramirez[46]). See ftp site for color image.

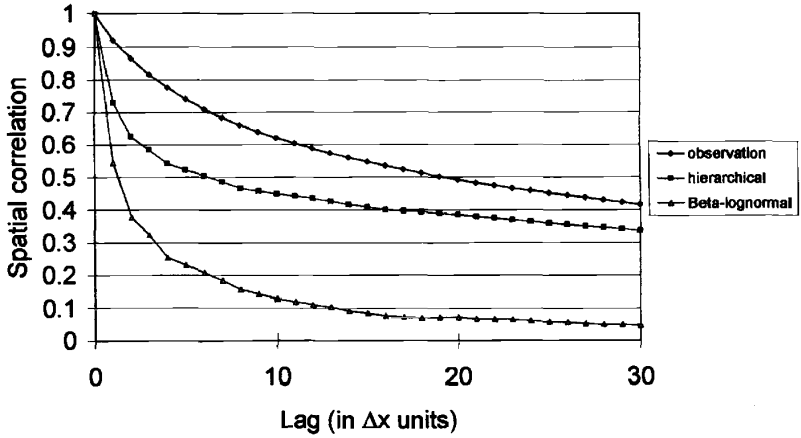


Figure 9 Comparison of correlation function for monthly rainfall (from Kang and Ramirez[46]). See ftp site for color image.

scale environment. Being able to parameterize the scaling characteristics of precipitation as a function of such variables is a prerequisite for implementing of downscaling methodologies based on random cascades. Perica and Foufoula-Georgiou[77] introduced a spatial downscaling scheme that is able to statistically reproduce the spatial heterogeneity of observed precipitation fields at subgrid scales while being conditioned on large-scale averages and physical properties. They computed multiscale standardized fluctuations using an orthogonal Haar wavelet decomposition and found that, at least for the range of scales of their analysis, these fluctuations exhibited normality and simple scaling. They also found that the scale-independent parameter H characterizing the simple scaling behavior of the standardized fluctuations was strongly dependent on the convective instability of the prestorm environment, namely on the convective available potential energy. The utility of the model in reproducing the small-scale statistical variability of precipitation as well as the fraction of area covered by rain at all subgrid scales was demonstrated,[77] and the relationship between H and the convective available potential energy of the prestorm

TABLE 2 Comparison of Basic Statistics for Observed (July 1997) and Simulated Precipitation Fields

	Observation	Nonparametric Superposition	Log-normal
Mean	0.296	0.295	0.300
Standard deviation	0.0257	0.0221	0.544
Skewness coefficient	1.174	1.185	16.78
Kurtosis	8.624	6.207	654.6

From Kang and Ramirez.[46]

environment established.[77] On the other hand, the relationship between the β-log-normal random cascade model parameters and the mean of the large-scale precipitation intensity was also observed and established.[46,75]

Most downscaling methodologies proposed in the literature only deal with the spatial variability of the precipitation field. The temporal evolution of the fields is usually described independently of the spatial downscaling, so that these schemes do not properly account for the temporal correlation structure, i.e., persistence, of the precipitation fields at subgrid scales. Recently, the linkage between the spatial and temporal scaling of precipitation fields has been explicitly addressed.[12,75,113,114] Over and Gupta[75] propose a model for space–time description of rainfall distributions based on multiplicative random cascades with independent weights in space, which are time varying according to an imposed structure. Carsteanu and Foufoula-Georgiou[12] argue that space and time variations of rainfall are necessarily connected. They postulate and experimentally verify a Taylor-like hypothesis stating that the power law variation for the moments is the same in time and space. Venugopal et al.[114] found that for spatial scales of 2 to 30 km and for temporal scales of 10 min to several hours, the evolution of precipitation remained statistically invariant under a transformation of the type $t \sim L^z$, where z is a so-called dynamic scaling exponent. That is, they found that the space–time organization of rainfall fields is scale-invariant and that its characteristics can be obtained by a simple renormalization of the space and time coordinates as implied by the $t \sim L^z$ transformation. They used the above results to develop a space–time precipitation downscaling scheme that is capable of preserving not only the spatial correlation of precipitation but also the temporal correlation at subgrid scales.

Finally, Seed et al.[99] have modeled the space–time behavior of radar precipitation using a multiplicative bounded (multifractal) cascade, each level of which was linked to the same level at the next time step via a different ARMA(1,1) model. Also Pegram and Clothier,[76] developed the so-called *string-of-beads* model in which power-law filtering of Gaussian random fields in space and time is used to capture the correlation structure of the rainfall process. Two autoregressive models, one at the image scale, the other at the pixel scale, drive the string-of-beads model. The spatial power-law filtering then ensures that the generated fields scale correctly in space and time.

REFERENCES

1. Aksoy, H., and M. Bayazit, A model for daily flows of intermittent streams, *Hydrol. Proc.*, *14*(10), 1725–1744, 2000.
2. Anthes, R. A., and T. T. Warner, Development of hydrodynamic models suitable for air pollution and other mesometeorological studies, *Monthly Weather Rev.*, *106*, 1045–1078, 1978.
3. Bartolini, P., and J. D. Salas, Modeling of streamflow processes at different time scales, *Water Resour. Res.*, *29*(8), 2573–2588, 1993.

4. Black, T. J., The new NMC mesoscale eta model: Description and forecast examples, *Weather Forecast.*, *9*, 265–278, 1994.
5. Box, G. E. P., and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, Holden-Day, San Francisco, 1976.
6. Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Addison-Wesley, Reading, MA, 1985.
7. Buishand, T. A., Some remarks on the use of daily rainfall models, *J. Hydrol.*, *36*, 295–308, 1977.
8. Burian, S. J., S. R. Durrans, S. Tomic, R. L. Pimentel, and C. N. Wai, Rainfall disaggregation using artificial neural networks, *ASCE J. Hydrol. Eng.*, *5*(3), 299–307, 2000.
9. Burlando, P., and R. Rosso, Comment on "Parameter estimation and sensitivity analysis for the modified Bartlett-Lewis Rectangular pulses model of rainfall by Islam et al., *J. Geophys. Res.*, vol. 95, no. D3, 1990, p. 2093–2100," *J. Geophys. Res.*, *96*(D5), 9391–9395, 1991.
10. Burlando, P., and R. Rosso, Stochastic models of temporal rainfall: Reproducibility, estimation and prediction of extreme events, in J. Marco Segura, R. Harboe, and J. D. Salas (Eds.), *Stochastic Hydrology and its Use in Water Resources Systems Simulation and Optimization*, Kluwer Academic Publishers, The Netherlands, 1993, pp. 137–173.
11. Cadavid, L. G., J. D. Salas, and D. C. Boes, Disaggregation of short-term precipitation records, in *Water Resources Papers*, Vol. 106, Colorado State University, Fort Collins, CO, 1992.
12. Carsteanu, A., and E. Foufoula-Georgiou, Assessing dependence among weights in a multiplicative cascade model of temporal rainfall, *J. Geophy. Res.*, *101*(D21), *26*, 363–26, 370, 1996.
13. Chang, T. J., M. L. Kavvas, and J. W. Delleur, Daily precipitation modeling by discrete autoregressive moving average processes, *Water Resour. Res.*, *20*, 565–580, 1984.
14. Chebaane, M., J. D. Salas, and D. C. Boes, Product periodic autoregressive processes for modeling intermittent monthly streamflows, *Water Resour. Res.*, *32*(5), 1513–1518, 1995.
15. Chin, E. H., Modeling daily precipitation occurrence process with Markov chain, *Water Resour. Res.*, *13*(6), 949–956, 1977.
16. Claps, P., F. Rossi, and C. Vitale, Conceptual-stochastic modeling of seasonal runoff using autoregressive moving average models and different scales of aggregation, *Water Resour. Res.*, *29*(8), 2545–2559, 1993.
17. Cowpertwait, P. S. P., and P. E. O'Connell, A regionalized Neyman-Scott model of rainfall with convective and stratiform cells, *Hydrol. Earth Syst. Sci.*, *1*, 71–80, 1997.
18. Delleur, J. W., and M. L. Kavvas, Stochastic models for monthly rainfall forecasting and synthetic generation, *J. Appl. Meteor.*, *17*, 1528–1536, 1978.
19. Dudhia, J., A nonhydrostatic version of the Penn State/NCAR mesoscale model: Validation tests and the simulation of an Atlantic cyclone and cold front, *Monthly Weather Rev.*, *121*, 1493–1513, 1993.
20. Eagleson, P., Climate, soil and vegetation, 2, The distribution of annual precipitation derived from observed storm sequences, *Water Resour. Res.*, *14*, 713–721, 1978.
21. Eltahir, E. A. B., A feedback mechanism in annual rainfall in Central Sudan, *J. Hydrol.*, *110*, 323–334, 1989.

22. Entekhabi, D., I. Rodriguez-Iturbe, and P. S. Eagleson, Probabilistic representation of the temporal rainfall process by a modified Neyman-Scott rectangular pulse model: Parameter estimation validation, *Water Resour. Res.*, *25*(2), 295–302, 1989.
23. Entekhabi, D., and P. S. Eagleson, Land surface hydrology parameterization for atmospheric general circulation models including subgrid scale spatial variability, *J. Climate*, *2*(8), 816–831, 1989.
24. Epstein, D., and J. A. Ramírez, Spatial disaggregation for studies of climatic hydrologic sensitivity, *ASCE J. Hydr. Div.*, *120*(12), 1449–1467, 1994.
25. Evora, N. D., and J. R. Rousselle, Hybrid stochastic model for daily flows simulation in semiarid climates, *ASCE J. Hydrol.*, *5*(1), 33–42, 2000.
26. Ewen, J., G. Parkin, and P. E. O'Connell, SHETRAN: Distributed River Basin flow and transport modeling system, *ASCE J. Hydrol. Eng.*, *5*(3), 250–258, 2000.
27. Fernandez, B., and J. D. Salas, Periodic gamma autoregressive processes for operational hydrology, *Water Resour. Res.*, *22*(10), 1385–1396, 1986.
28. Fernandez, B., and J. D. Salas, Gamma-autoregressive models for streamflow simulation, *J. Hydr. Eng. ASCE*, *116*(11), 1403–1414, 1990.
29. Fiering, M. B., and B. B. Jackson, *Synthetic Streamflows*, Water Resources Monograph 1, American Geophysical Union (AGU), Washington, DC, 1971.
30. Foufoula-Georgiou, E., and P. Guttorp, Compatibility of continuous rainfall occurrence models with discrete rainfall observations, *Water Resour. Res.*, *22*, 1316–1322, 1986.
31. Foufoula-Georgiou, E., and D. P. Lettenmaier, A Markov renewal model of rainfall occurrences, *Water Resour. Res.*, *23*(5), 875–884, 1987.
32. Foufoula-Georgiou, E., and W. Krajewski, Recent advances in rainfall modelling, estimation and forecasting, *Rev. Geophys.*, Suppl., 1125–1137, July 1995.
33. Giorgi, F., and L. O. Mearns, Approaches to the simulation of regional climate change: A review, *Rev. Geophys.*, *29*(2), 191–216, 1991.
34. Giorgi, F., M. R. Marinucci, and G. T. Bates, Development of a second-generation regional climate model (RegCM2). Part I: Boundary-layer and radiative transfer processes, *Monthly Weather Rev.*, *121*, 2794–2813, 1993a.
35. Giorgi, F., M. R. Marinucci, and G. T. Bates, Development of a second-generation regional climate model (RegCM2). Part II: Convective processes and assimilation of lateral boundary conditions, *Monthly Weather Rev.*, *121*, 2814–2832, 1993b.
36. Grygier, J. C., and J. R. Stedinger, Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res.*, *24*, 1574–1584, 1988.
37. Gupta, V. K., and E. Waymire, Multiscaling properties of spatial rainfall and river flow distributions, *J. Geoph. Res.*, *95*(D3), 1999–2009, 1990.
38. Gupta, V. K., and E. Waymire, A statistical analysis of mesoscale rainfall as a random cascade, *J. Appl. Meteor.*, *12*(2), 251–267, 1993.
39. Guttorp, P., *Stochastic Modeling of Scientific Data*, Chapman Hall, London, 1995.
40. Gyasi-Agyei, Y., and G. R. Willgoose, A hybrid model for point rainfall modeling, *Water Resour. Res.*, *33*(7), 1699–1706, 1997.
41. Hershenhorn, J., and D. A. Woolhiser, Disaggregation of daily rainfall, *J. Hydrol.*, *95*, 299–322, 1987.
42. Hipel, K. W., and A. I. McLeod, *Time Series Modeling of Water Resources and Environmental Systems*, Elsevier, Amsterdam, 1994.

43. Hirsch, R. M., Synthetic hydrology and water supply reliability, *Water Resour. Res.*, *15*(6), 1603–1615, 1979.
44. Hosking, J. R. M., Fractional differencing, *Biometrika*, *68*, 165–176, 1981.
45. Intergovernmental Panel on Climate Change (IPCC), *Summary for Policymakers*, report of Working Group I of the IPCC, available on-line, http://www.ipcc.ch/, 2001.
46. Kang, B., and J. A. Ramirez, Comparative study of the statistical features of random cascade models for spatial rainfall downscaling, in J. A. Ramirez (Ed.), *Proc. AGU Hydrol. Days 2001*, Hydrology Days Publications, Fort Collins, CO, 2001, pp. 151–164.
47. Karl, T. R., W. C. Wang, M. E. Schlesinger, R. W. Knight, and D. Portman, A method of relating general circulation model simulated climate to the observed local climate. Part I: Seasonal statistics. *J. Climate*, *3*, 1053–1079, 1990.
48. Katz, R. W., On some criteria for estimating the order of a Markov chain, *Technometrics*, *23*(3), 243–249, 1981.
49. Katz, R. W., and M. B. Parlange, Generalizations of chain-dependent processes: Application to hourly precipitation, *Water Resour. Res.*, *31*, 1331–1341, 1995.
50. Kavvas, M. L., L. J. Cote, and J. W. Delleur, Time resolution of the hydrologic time series models, *J. Hydrol.*, *32*, 347–361, 1977.
51. Kavvas, M. L., and J. W. Delleur, A stochastic cluster model of daily rainfall sequences, *Water Resour. Res.*, *17*(4), 1151–1160, 1981.
52. Kelman, J, A stochastic model for daily streamflow, *J. Hydrol.*, *47*, 235–249, 1980.
53. Koch, R. W., A stochastic streamflow model based on physical principles, *Water Resour. Res.*, *21*(4), 545–553, 1985.
54. Koepsell, R. W., and J. B. Valdes, Multidimensional rainfall parameter estimation from a sparse network, *ASCE J. Hydr. Eng.*, *117*(7), 832–850, 1991.
55. Krajewski, W. F., and J. A. Smith, Sampling properties of parameter estimators for a storm field rainfall model, *Water Resour. Res.*, *25*(9), 2067–2075, 1989.
56. Lane, W. L., *Applied Stochastic Techniques (Last Computer Package), User Manual*, Division of Planning Tech. Services, Bureau of Reclamation, Denver, CO, 1979.
57. Lane, W. L., Corrected parameters estimates for disaggregation schemes, in V. P. Singh (Ed.), *Statistical Analysis of Rainfall and Runoff*, Water Resources Publications (WRP), Littleton, CO, 1982.
58. Lanza, L. G., A conditional simulation Model of intermittent rain fields, *Hydrol. Earth Sys. Sci. 4*(1), 173–183, 2000.
59. Leavesley, G. H., R. W. Lichty, B. M. Troutman, and L. G. Saindon, *Precipitation-Runoff-Modelling-System—User's Manual*, USGS Water Resour. Invest. Report, U.S. Geological Survey, 83-4238, 1983.
60. Le Cam, L. A., A stochastic description of precipitation, in J. Newman (Ed.), *Proc. IV Berkeley Symp. on Math., Statis. & Prob.*, University of Calif. Press, Berkeley, 1961, pp. 165–186.
61. Lettenmaier, D. P., and S. J. Burges, Operational assessment of hydrologic models of long-term persistence, *Water Resour. Res.*, *13*(1), 113–124, 1977.
62. Loucks, D. P., J. R., Stedinger, and D. Haith, *Water Resources Systems Planning and Analysis*, Prentice Hall, Englewood Clifts, NJ, 1981.
63. Lovejoy, S., and B. B. Mandelbrot, Fractal properties of rain and a fractal model, *Tellus*, *37A*, 209–232, 1985.

64. Mandelbrot, B. B., and J. R. Wallis, Computer experiments with fractional Gaussian noises: Part 1, Averages and variances, *Water Resour. Res.*, *5*(1), 228–241, 1969.

65. Matalas, N. C., Mathematical assessment of synthetic hydrology, *Water Resour. Res.*, *3*(4), 937–945, 1967.

66. McKerchar, A. I. and J. W. Delleur, Application of seasonal parametric stochastic models to monthly flow data, *Water Resour. Res.*, *10*, 246–255, 1974.

67. Mellor, D., The modified turning bands (MTB) model for space-time rainfall. I. Model definition and properties, *J. Hydrol.*, *175*(1–4), 113–127, 1996.

68. Murrone, F., F. Rossi, and P. Claps, Conceptually-based shot noise modeling of streamflows at short time interval, *Stochast Hydrol. Hydraul.*, *11*(6), 483–510, 1997.

69. Neyman, J., and E. L. Scott, Statistical approach to problems of cosmology, *J. R. Stat. Soc. Ser. B*, *20*(1), 1–43, 1958.

70. Obeysekera, J. T. B., and J. D. Salas, Modeling of aggregated hydrologic series, *J. Hydrol.*, *86*, 197–219, 1986.

71. Obeysekera, J. T. B., G. Tabios, and J. D. Salas, On parameter estimation of temporal rainfall models, *Water Resour. Res.*, *23*(10), 1837–1850, 1987.

72. O'Connell, P. E., A simple stochastic modeling of Hurst's law, in *1971 Warsaw Symp. in Mathematical Models in Hydrology*, International Association of Hydrologic Sciences, Pub. vol. 100, No. 1, 1974, pp. 169–187.

73. O'Connell, P. E. Stochastic modeling of long-term persistence in streamflow sequences, Ph.D. dissertation, Imperial College of Science and Technology, University of London, England, 1974.

74. Ormsbee, L. E., Rainfall disaggregation model for continuous hydrologic modeling, *ASCE J. Hydraul. Eng.*, *115*(94), 507–525, 1989.

75. Over, T. M., and V. J. Gupta, A space-time theory of mesoscale rainfall using random cascades, *J. Geophys. Res.*, *101*(D21), 26319–26331, 1996.

76. Pegram, G. G. S., and A. N. Clothier, High resolution space-time modeling of rainfall: The string of beads model, WRC Report no. 752/1/99, report to the Water Research Commission, Pretoria, South Africa, 1999.

77. Perica, S. E., and E. Foufoula-Georgiou, Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions, *J. Geophys. Res. Atmos. 101*(D21), 26347–26361, 1996.

78. Pielke, R. A., and R. Avissar, Influence of landscape structure on local and regional climate, *Landscape Ecol.*, *4*, 133–155, 1990.

79. Pielke, R. A., W. R. Cotton, R. L. Walko, C. J. Tremback, M. E. Nicholls, M. D. Moran, D. A. Wesley, T. J. Lee, and J. H. Copeland, A comprehensive meteorological modeling system—RAMS, *Meteor. Atmos. Phys.*, *49*, 69–91, 1992.

80. Pielke, Sr., R. A., Overlooked issues in the U.S. national climate and IPCC assessments, Preprints, in *11th Symp. on Global Change Studies, 80th AMS Annual Meeting*, Long Beach, CA, January 9–14, 2000, pp. 32–35.

81. Pielke, Sr., R. A., and L. Guenni, Vulnerability assessment of water resources to changing environmental conditions, *IGBP Newslett.*, *39*, 21–23, 1999.

82. Ramirez, J. A., and R. L. Bras, Conditional distributions of Neyman-Scott models for storm arrivals and their use in irrigation control, *Water Resour. Res.*, *21*, 317–330, 1985.

83. Ramírez, J. A., and S. Senarath, A statistical-Dynamical parameterization of canopy interception and land surface-atmosphere interactions, *J. Climate*, *13*, 4050–4063, 2000.

84. Rasmussen, P. F., J. D. Salas, L. Fagherazzi, J. C. Rassam, and B. Bobee, Estimation and validation of contemporaneous PARMA models for streamflow simulation, *Water Resour. Res.*, *32*(10), 3151–3160, 1996.

85. Richardson, C. W., and D. A. Wright, *WGEN: A Model for Generating Daily Weather Variables*, U.S. Department of Agriculture, Agriculture Research Service, ARS-8, August, 1984.

86. Rodriguez–Iturbe, I., V. K. Gupta, and E. Waymire, Scale considerations in the modeling of temporal rainfall, *Water Resour. Res.*, *20*(11), 1611–1619, 1984.

87. Rodriguez–Iturbe, I., D. R. Cox, and V. Isham, Some models for rainfall based on stochastic point processes, *Proc. R. Soc. Lond. Ser. A*, *410*, 269–288, 1987.

88. Rodriguez–Iturbe, I., B. Febres de Power, and J. B. Valdes, Rectangular pulses point process models for rainfall: Analysis of empirical data, *J. Geophys. Res.*, *92*(D8), 9645–9656
1987.

89. Rodriguez–Iturbe, I., B. Febres de Power, M. B. Sharifi, and K. Georgakakos, Chaos in rainfall, *Water Resour. Res.*, *25*(7), 1667–1675, 1989.

90. Roldan, J., and D. A. Woolhiser, Stochastic daily precipitation models: 1. A comparison of occurrence processes, *Water Resour. Res.*, *18*(5), 1451–1459, 1982.

91. Salas, J. D., D. C. Boes, V. Yevjevich, and G. G. S. Pegram, Hurst phenomenon as a pre-asymptotic behavior, *J. Hydrol.*, *44*(1), 1–15, 1979.

92. Salas, J. D., J. R. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, Water Resources Publications, Littleton, CO, 1980.

93. Salas, J. D., and D. C. Boes, Shifting level modelling of hydrologic series, *Adv. Water Resour.*, *3*, 59–63, 1980.

94. Salas, J. D., and M. Chebaane, Stochastic modeling of monthly flows in streams of arid regions, in *Proc. Intern. Symp. HY&IR Div. ASCE*, San Diego, CA, 1990, pp. 749–755.

95. Salas, J. D., and M. W. Abdelmohsen, Determining streamflow drought statistics by stochastic simulation, in *Proc. U.S.-PRC Bilateral Symp. on Droughts and Arid-Region Hydrology*, Tucson, AZ, September 16–20, 1991, U.S. Geological Survey Open–File Report No. 91–244, 1991.

96. Salas, J. D., and J. T. B. Obeysekera, Conceptual basis of seasonal streamflow time series models, *ASCE J. Hydraul. Eng.*, *118*(8), 1186–1194, 1992.

97. Salas, J. D., Analysis and modeling of hydrologic time series, in D. R. Maidment (Ed.) *Handbook of Hydrology*, McGraw-Hill Book, New York, 1993, Chapter 19.

98. Santos, E., and J. D. Salas, Stepwise disaggregation scheme for synthetic hydrology, *ASCE J. Hydr. Eng.*, *118*(5), 765–784, 1992.

99. Seed, AW, R Srikanthan, and M. Menabde, A space and time model for design storm rainfall, *J. Geophy. Res.*, accepted for publication.

100. Sharma, A., D. G. Tarboton, and U. Lall, Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, *33*(2), 291–308, 1997.

101. She, Z. S., and E. C. Waymire, Quantized energy cascade and log-Poisson statistics in fully developed turbulence, *Phys. Rev. Lett.*, *74*(2), 1995.

102. Shukla, J., J. Anderson, D. Baumherner, C. Brankovic, Y. Chang, E. Kalnay, L. Marx, T. Palmer, D. Paolino, J. Ploshay, S. Schubert, D. Straus, M. Suarez, and J. Tribbia, Dynamical seasonal prediction, *Bull. Am. Meteor. Soc*, *81*, 2593–2606, 2000.

103. Smith, J. A., and A. F. Karr, A point process model of summer season rainfall occurrences, *Water Resour. Res.*, *19*(1), 95–103, 1983.

104. Smith, J. A., and W. F. Krajewski, Statistical modeling of space–time rainfall using radar and rain gage observations, *Water Resour. Res.*, *23*(10), 1893–1900, 1987.

105. Stedinger, J. R., and R. M. Vogel, Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res.*, *20*(1), 47–56, 1984.

106. Stedinger, J. R., D. P. Lettenmaier, and R. M. Vogel, Multisite ARMA(1,1) and disaggregation models for annual streamflow generation, *Water Resour. Res.*, *21*, 497–509, 1985a.

107. Stedinger, J. R., D. Pei, and T. A. Cohn, A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res.*, *21*(5), 665–675, 1985b.

108. Tarboton, D. G., A. Sharma, and U. Lall, Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res.*, *34*, 107–119, 1998.

109. Tessier, Y., S. Lovejoy, and D. Schertzer, Universal multifractals: Theory and observations for rain and clouds, *J. Appl. Meteorol.*, *32*(2), 223–250, 1993.

110. Treiber, B., and E. J. Plate, A stochastic model for the simulation of daily flows, *Hydrol. Sci. Bull.*, *22*(1), 175–192, 1977.

111. Valencia, D. R., and J. C. Schaake, Jr., Disaggregation processes in stochastic hydrology, *Water Resour. Res.*, *9*(3), 580–585, 1973.

112. Vecchia, A., J. T. B. Obeysekera, J. D. Salas, and D. C. Boes, Aggregation and estimation for low-order periodic ARMA models, *Water Resour. Res.*, *19*(5), 1297–1306, 1983.

113. Venugopal, V., and E. Foufoula-Georgiou, Energy decomposition of rainfall in the time-frequency-scale domain using wavelet packets, *J. Hydrol.*, *187*, 3–27, 1996.

114. Venugopal, V., E. Foufoula-Georgiou, and V. Sapozhnikov, Evidence of dynamic scaling in space-time rainfall, *J. Geophys. Res.*, *104*(D24), 31599–31610, 1999.

115. Walko, R. L., L. Band, J. Baron, T. G. Kittel, R. Lammers, T. J. Lee, D. Ojima, R. A. Pielke, Sr., C. Taylor, C. Tague, C. J. Tremback, and P. L. Vidale, Coupled atmosphere-biophysics-hydrology models for environmental modeling, *J. Appl. Meteor.*, *39*, 931–944, 2000.

116. Waymire, E., V. K. Gupta, and I. Rodriguez-Iturbe, A spectral theory of rainfall intensity at the meso-β scale, *Water Resour. Res.*, *20*(10), 1453–1465, 1984.

117. Wigley, T. M. L., P. D. Jones, K. R. Briffa, and G. Smith, Obtaining sub-grid-scale information from coarse resolution general circulation model output, *J. Geophys. Res.*, *95*(D2), 1943–1953, 1990.

118. Wilks, D. S., *Statistical Methods in the Atmospheric Sciences*, Academic, San Diego, CA, 1995.

119. Woolhiser, D. A., and H. B. Osborn, A stochastic model of dimensionless thunderstorm rainfall, *Water Resour. Res.*, *21*(4), 511–522, 1985.

120. Yevjevich, V., *Stochastic Processes in Hydrology*, Water Resouces Publications, Littleton, CO, 1972.

CHAPTER 34

STOCHASTIC FORECASTING OF PRECIPITATION AND STREAMFLOW PROCESSES

JUAN B. VALDÉS, PAOLO BURLANDO, AND JOSÉ D. SALAS

1 INTRODUCTION

Over the past two decades, considerable research has been carried out in hydrology on developing mathematical tools and approaches for short- and long-term precipitation and streamflow forecasting. The forecasts may be concerned with flood warning, flood control, water quality control, navigation, energy production, and irrigation. Hydrologic *forecasting* signifies estimating the time of occurrence and the magnitude of a hydrological event before its actual occurrence (e.g., estimating daily streamflow with days or weeks in advance), i.e., an estimate of the future states of the hydrological phenomena is obtained in *real-time*. The adjective real-time is often used to reinforce the distinction between forecasting (the estimation of future hydrologic events based on the currently available data) and simulation, sometimes called long-term prediction (the estimation of equally likely scenarios of hydrologic events without necessarily conditioning on real-time data). In short, forecasting is generally used for operational and management purposes while simulation is used for design and planning purposes.

Forecasting of hydrological processes is an important tool for many water resources management and operational problems. For example, rainfall and streamflow forecasting hours, days, weeks, and months in advance (depending on the particular case at hand) are important for many flood warning, evacuation, and mitigation plans and actions. The U.S. National Weather Service (NWS) routinely issues precipitation forecasts (throughout the year) for all the U.S. territories and flow forecasts at key control points of the stream network systems in the United

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

States. Forecasting the number of hurricanes of certain strengths that may occur in the following year (Gray et al., 1994) and forecasting the path and the intensity of an ongoing hurricane, have been regular activities of the National Oceanic and Atmospheric Administration's (NOAA's) Hurricane Center. From the hydrologic and water resources perspectives, forecasting hurricanes has many implications, particularly as they relate to the occurrence of floods. In systems involving reservoirs, hurricane forecasts are useful for planning and implementing special operating rules to cope with impending floods. In small river systems that may be subject to flash floods, forecasting rainfall and streamflow a few hours in advance may be critical for implementing emergency actions such as alerting and warning the public. On the other hand, in large systems, such as the Mississippi River in the United States or the Paraná River in Argentina, flood occurrences may develop through several weeks and months. In these cases rainfall and flow forecasts are usually needed with lead times of weeks and months. Also in river systems where spring and summer runoff occurs from snowmelt, forecasts are usually needed weeks and months in advance for planning water supply and hydropower systems operations and for preparing for possible snowmelt floods. In such cases, determining the current amount of snow pack in the system and snow properties is of outmost importance. The development of reservoir operating rules and the real-time operation of reservoir systems may require hourly, daily, weekly, monthly, and yearly forecasts depending of the particular case at hand. Forecasts of rainfall, snowfall, snow pack, soil moisture, evaporation, streamflow, reservoir levels, river levels, and groundwater heads are generally needed in most cases of practical interest.

Forecasting of hydrologic processes has been developed using similar approaches as for simulation, although many models and techniques are unique either for simulation or forecasting. This chapter emphasizes forecasting based on stochastic and probabilistic techniques. Also, the emphasis will be on precipitation and streamflow processes, although many of the methods and models included herein are equally applicable for other hydroclimatic processes as well as evapotranspiration, soil moisture, surface and groundwater levels, and sea surface temperature.

In developing precipitation and streamflow forecasting models, one must be aware of the large uncertainty in the model parameters because of inadequate historical data of the relevant processes under consideration. Furthermore the model parameters may be expected to change slowly/rapidly with time, but the exact nature of the change is not predictable. In such cases, it is highly desirable to develop a model that has self-learning capabilities, so that it can adapt itself to the current situation (Brown and Hwang, 1997). For this purpose, filters have been formulated in the literature under the assumption that dynamic system parameters and input/measurement error statistics are known. This is not the case for precipitation and streamflow forecasting and additional estimation techniques are necessary. The sequential estimation procedure, known as the *Kalman filter*, is optimal under such conditions. However, if the actual values of system coefficients and covariances are different from those used in state estimation, then the filter is suboptimal: State estimates may contain more errors than is necessary and, in some cases, diverge from the neighborhood of the true values. State estimates could be improved by

simultaneously estimating the uncertain parameters and the statistics. This additional information may be used to adapt the filter gains and model coefficients to the measurements. Adaptive filters may perform as well as optimal filters in the limit (Stengel, 1986).

Nonstationary characteristics are conventionally assumed to arise from the presence of one or more integrators in the stochastic part of the signal generation process. This applies in those cases where the model of the underlying time series data can only be characterized adequately by parameters, which vary over time in some significant manner. In all these situations the Kalman filter provides information on the possible nature of these parametric variations. Other statistical tools that are used for short- and long-term forecasting of precipitation and streamflows include methods based on regression models, autoregressive integrated moving average (ARIMA) models, ARMAX models, transfer function noise (TFN) models, and models based in artificial neural networks (ANN). In the next section a brief description of the Kalman filter will be made because of the ample use of this technique in hydroclimatic forecasting and because many of the above models can be used in conjunction with the Kalman filter. Subsequent sections will include many of the referred models and techniques for precipitation and streamflow forecasting.

2 ADAPTIVE PREDICTION: THE KALMAN FILTER

Since its introduction the Kalman filter has become a powerful tool in the fields of estimation and control theory (Kalman, 1960; Kalman and Bucy, 1961). As systems become more complex and noise becomes present in both input and output variables, it is then necessary to search for statistical solutions that can take advantage of past performance and adjust future forecasts accordingly. It is viewed as a complementary tool to the mathematical modeling of the rainfall–runoff process rather than a substitute because the knowledge of the underlying mechanism of the hydrologic process is essential for a successful implementation of the filter. The main purpose of this section is to present an introductory view of the Kalman filter rather than a thorough theoretical explanation of the statistical properties of the filter. For a successful application of the filter to real-time forecasting of hydroclimatic variables, the main hypothesis and limitations of the filter must be understood.

There are three different types of estimation problems depending on how the observations are used:

- *Filtering* The observations $\mathbf{z} = \{z_1\ z_2, \ldots, z_t\}$ are used for filtering to obtain an estimate $\mathbf{x}_{t|t}$ of the state of the system $\mathbf{x}_t$.
- *Smoothing* The observations $\mathbf{z} = \{z_1\ z_2, \ldots, z_t\ z_{t+1}\}$ are used for smoothing to obtain an estimate $\mathbf{x}_{t|t+1}$ of the state of the system $\mathbf{x}_t$.
- *Prediction* The observations $\mathbf{z} = \{z_1\ z_2, \ldots, z_{t-1}\}$ are used in prediction to obtain an estimate $\mathbf{x}_{t|t-1}$ of the state of the system $\mathbf{x}_t$.

For a detailed discussion on the topic the reader is referred to specialized books (e.g., Brown and Hwang, 1996). This reference also includes the software for some applications. Recursive algorithms are ideal for estimation of time-varying parameters. Modifications based on stochastic modeling of the parameter variations lead naturally to the development of the Kalman filter and the estimation of time-varying states in stochastic dynamic systems. Kalman considerably extended the state-estimation and filter theory of time-varying parameters or states so as to handle the analysis of nonstationary time series and provide a natural approach to the analysis of time series data that are assumed to be generated from stochastic state-space equations.

When modeling a system that evolves through time, specifically a stochastic process that is defined in discrete time, one would like to put the system in a *state-space* form or in the so-called *state of the system* vector $\mathbf{x}_t$. (Most linear models can be put into state-space form; nonlinear models can be linearized by using Taylor series expansion to reformulate them in state-space form.) If future values of the state of the system, $\mathbf{x}_{t+s}$, $s = 1, 2, \ldots$, can be modeled using knowledge of x_t (i.e., $\mathbf{x}_t$ contains all the required information about the previous values $\mathbf{x}_{t-s}$, $s = 1, 2, \ldots$), we obtain what is called a Markovian system. The best description of $\mathbf{x}_t$ using $\mathbf{x}_{t-1}$, $\mathbf{x}_{t-2}, \ldots$ can be modeled as

$$\mathbf{x}_t = \Phi(\mathbf{x}_{t-1},\ t-1) + \Gamma(\mathbf{w}_t,\ t) \tag{1}$$

Equation (1) is called the *state equation* of the system, where $\Phi(\cdot)$ is the transition function, $\Gamma(\cdot)$ is the noise transition function, $\mathbf{w}_t$ is the vector of system noises that describes the part of $\mathbf{x}_t$ that is not explained by $\mathbf{x}_s$, $s < t$, and is assumed independent of $\mathbf{x}_s$ and $\mathbf{w}_s$ for $s < t$. When $\Phi(\cdot)$ and $\Gamma(\cdot)$ do not vary with time dependence, the system is referred to as *stationary*.

In most applications, the state of the system, $\mathbf{x}_t$, is not directly observed but rather measured in an observation vector $\mathbf{z}_t$, which is a function of $\mathbf{x}_t$, corrupted by measurement noise $\mathbf{v}_t$. This may be written as:

$$\mathbf{z}_t = \mathbf{H}_t(\mathbf{x}_t,\ t) + \mathbf{v}_t \tag{2}$$

This equation is called the *observation equation of the filter*, and Eqs. (1) and (2) together constitute the heart of the Kalman filter; they may represent linear or nonlinear systems. The filtering problem is to estimate $\mathbf{x}_t$ from the observations $\mathbf{z}_1, \ldots, \mathbf{z}_t$, which are corrupted by measurement noises. If both the system and the observations are assumed linear, Eqs. (1) and (2) will have the following form:

$$\mathbf{x}_t = \Phi_t \mathbf{x}_{t-1} + \Gamma_t \mathbf{w}_t \tag{3}$$

$$\mathbf{z}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{v}_t \tag{4}$$

where Φ_t, Γ_t, and $\mathbf{H}_t$ are known matrices.

The stochastic properties of the system and measurement noises have to be defined in order to apply the Kalman filter. Also the properties of the initial state

of the system have to be defined. The main challenges in the application of the filter to forecasting of precipitation and streamflow are to define the appropriate matrices and other components of the filter in terms of the hydrologic variables and to estimate them from available data and knowledge of the physical process. In subsequent sections the application of various statistical techniques including the Kalman filter for forecasting hydroclimatic processes, particularly precipitation and streamflow, is presented.

3 STOCHASTIC PRECIPITATION FORECASTING

Precipitation forecasting is of great significance for water resources management and flood protection, although it is not an easy task. Rainfall forecasts based on the analysis of the temporal and spatial evolution of the meteorological phenomena would be always desirable. Considerable progress has been made in this respect by using numerical weather prediction approaches and general circulation models. However, information from such sources is not always available in operational form. In this situation rainfall forecasts can be made based on the persistence characteristics of current and past rainfall measurements, even though the accuracy of such forecasts may suffer because of the lack of the physical aspects involved in the precipitation phenomena. A number of examples of precipitation forecasting by statistical, stochastic, or probabilistic techniques can be found in the literature. They include regression techniques, Markov chains, ARMA-type methods, probability-function-based approaches, and artificial neural networks (ANNs). All of these approaches have been used for short-term and for mid- and long-term forecasting, as illustrated below.

Short-Term Forecasting

Quantitative precipitation forecasting, often denoted as QPF, is one of the major tasks in flood forecasting. It has been demonstrated that QPF allows extending the lead time of flood forecasts and improving the accuracy of flood estimates for a given forecast lead time (Brath et al., 1988). Although research in the field of numerical weather prediction has achieved significant progress in recent years (see, e.g., Bougeault et al., 2000), forecasting techniques based on stochastic and statistical modeling are useful especially for operational purposes and in the context of mesoscale basins, which are characterized by rapid response time. However, because of the complexity of the rainfall phenomena, which exhibits significant spatial and temporal variability, nonstationarity, and nonlinearity, especially on small scales, rainfall forecasting by stochastic approaches involves a challenging feat and experience.

Early attempts to forecast rainfall were formulated as statistical black-box models used for storm tracking. For instance, Phanartzis (1979) developed a simple model for forecasting the direction of storm movement based on the cross-correlation of rainfall measured at a network of rain gages. A similar approach was developed by

Nguyen et al. (1978) to be used with radar storm tracking signals. A more sophisticated storm tracking statistical procedure based on Kalman filter was proposed by Johnson and Bras (1980). Also, French and Krajewski (1994) and French et al. (1994) used the Kalman filter for state updating and incorporation of uncertainty in a two-dimensional physically based model and surface meteorological observations. Furthermore, Sugimoto et al. (2001) also used the extended Kalman filter as a state estimator to update the model parameter of the conceptual model with new radar data and with forecasts from a numerical weather prediction model.

Other authors, such as Lardet and Obled (1994), generated scenarios of rainfall duration and volume by probability functions conditioned on past rainfall. Statistical methods based on classification trees were also used for QPF (Carter and Elsner, 1997; Carter et al., 2000). In other applications physically based model structures are combined with stochastic components to account for the uncertainties associated with model hypotheses and structure (Jinno et al., 1993). Kawamura et al. (1996, 1997) added a Gaussian white noise in time and space to an advection-diffusion model of space–time rainfall, to consider a certain degree of error and uncertainty inherent in rainfall modeling.

Other approaches try to overcome the intrinsic limitation of persistence-based methods for predicting rainfall, due to the short decorrelation time of the precipitation process, which has been shown to be of the order of approximately 20 min (Zawadzki, 1987). Four stochastically based approaches for forecasting short-term precipitation are presented below.

Point Process Models. The models based on point processes perform satisfactorily with respect to reproducing the cluster dependence properties of observed rainfall (Entekhabi et al., 1989) and related extreme properties (Burlando and Rosso, 1993). However, the formulation required for real-time forecasting is very complex. Ramirez and Bras (1985) developed an algorithm for forecasting storm arrivals assuming the Neyman–Scott white-noise model as the underlying rainfall-generating mechanism. They derived the general expressions for the distribution functions of the time to the next storm event, conditioned on part of the immediate rainfall history, and applied the algorithm for irrigation scheduling. French et al. (1992a) developed a real-time forecasting scheme based on the space–time model of Rodriguez-Iturbe and Eagleson (1987). The forecasting model consists of a single distributed state-space equation, which is used to derive the conditional mean and the conditional covariance of rainfall intensity. Updating of the rainfall field in real time is carried out by representing the model structure as a distributed parameter Kalman filter. While some work has been done in using point and cluster processes for real-time forecasting of precipitation, their development has been limited to research studies.

Regression-Based Methods. A good example of how rainfall forecasting based on statistical methods is useful for operational purposes is the U.S. National Weather Service's centralized statistical quantitative precipitation forecasts (Antolik, 2000). The statistical forecast is based on multiple linear regression (Glahn and

Lowry, 1972; Lowry and Glahn, 1976), where the rainfall amount over a given time interval is predicted as a function of meteorological variables, both observed and computed by numerical weather models. Despite the relative simplicity of the model, it often outperforms physically based methods and more complex techniques, depending on the proper identification of the predictors. The use of regression methods though is more common in long-term forecasting.

Markov Chains Approach. The theory of Markov chains has been suggested for short- and long-term forecasting of rainfall. For example, Bertoni et al. (1992) used a first-order Markov chain for real-time forecasting of rainfall for a few hours lead time, which in turn was used for flood forecasting. Historical rainfall data were classified in states that divide the range of rainfall variation into sequences of nonoverlapping intervals. The transition probabilities were estimated as $p_{ij} = n_{ij} / \sum_{j=1}^{r} n_{ij}$, $(i, j = 1, \ldots, r)$, where r is the number of states, and n_{ij} is the number of transitions from state i to j, which is computed from historical observations on a seasonal basis. The p_{ij} values are elements of the transition probability matrix, which is then used to estimate (forecast) the m-step (ahead) transition probability $p_{ij}^{(m)}$ on the basis of the incoming observations (i.e., the present state) and the given conditional nonexceedence probability. The selection of an appropriate nonexceeding probability is key in achieving acceptable rainfall forecasts. Yu and Yang (1997) adopted a similar approach and further analyzed the role played by the choice of the nonexceeding probability with respect to forecast accuracy. In addition to seasonal dependence, the nonexceeding probability strongly depends on storm profile, being considerably different in the raising limb than in the recession limb of the hyetographs.

Dahale and Puranik (2000) applied a six-state simple Markov chain to forecast 5-day spatial rainfall persistence of summer monsoons over the Indian region. Fraedrich and Müller (1983) used a five-state simple Markov chain, and Miller and Leslie (1984) adopted a four-state second-order model to predict rainfall probabilities from past weather states. One must note that high forecast skills are generally obtained for short lead times, and they significantly decrease with increasing lead times. Johnson and Bras (1980) combined forecasts of the mean rainfall rate throughout the event at each gage with the modeling of a random residual component based on a Markovian model. The choice of the optimal order of a Markov chain also plays a role in forecast accuracy. Akaike information criterion and Bayes information criterion can be used for this purpose (e.g., Tong, 1975; Katz, 1981; Gregory et al., 1992).

ARMA Models. Trotta et al. (1977), and Labadie et al. (1981) showed that ARMA and transfer function models can be used for modeling rainfall persistence. They used an autoregressive transfer function model for short-term rainfall forecasting for the purpose of improving the control of a sewer system. The model uses parameters estimated from historical data at the beginning of the storm event, when information of the ongoing event is still poor. As the storm progresses, the parameters are progressively tuned to reflect the increasing real-time information. This is done in

the estimation step by including weighting factors in a least-squares algorithm to account differently for the historical information and current rainfall event information. Obeysekera et al. (1987) showed that certain point process models widely applied for modeling short-term rainfall, such as the Poisson rectangular pulse (PRP) and the Neyman Scott rectangular pulse (NSRP), possess correlation structures like those of ARMA(1,1) and ARMA(2,2) models, respectively. Thus, in principle, ARMA models could be used for simulation and forecasting of short-term rainfall processes. Because ARMA models are stationary, and the underlying variable is normally distributed, their application to real-time forecasting of short-term precipitation, such as hourly and daily rainfall, requires certain procedures to be followed to take into account such requirements. Burlando et al. (1993) used the ARMA(2,2) model given as

$$Z_t = \sum_{j=1}^{2} \phi_j Z_{t-j} + \varepsilon_t - \sum_{j=1}^{2} \theta_j \varepsilon_{t-j} \tag{5}$$

where $Z_t = X_t - \mu$, X_t represents hourly rainfall, μ is the mean of X_t, ϕ_j and θ_j are the autoregressive and moving average coefficients, respectively, and ε_t is a normally distributed noise with mean zero and variance σ_ε^2.

Nonstationarity of the rainfall throughout the year was accounted for either by seasonal estimation of parameters based on the analysis of the continuous data set or by event-based parameter estimation carried out only on extracted nonzero rainfall events. In the latter approach a different parameter set was determined for each storm event considered. To account for the nonlinearity that characterizes storm precipitation events, some modifications were necessary for the estimation of the ARMA model (5) as shown schematically in Figure 1. Specifically, data are first transformed to account for non-normality by means of the Box–Cox transformation (Box and Cox, 1964), and the estimation of the model parameters is performed by an iterative adaptive least-squares technique. Thus, the data used for estimation are only those available as the storm event evolves through time, implicitly assuming a local stationarity. While the results based on the continuous data set were not satisfactory, the event-based application provided satisfactory results. Figure 2 shows an example of the forecast accuracy. A noticeable problem, however, is the one-hour phase shift that characterizes most of the forecasted events. Toth et al. (2000) obtained similar results by slightly modifying the procedure introduced by Burlando et al. (1993), in that the event-based parameter estimation was carried out on the basis of a moving window of fixed length rather than on the complete event data sequence. Transformation of data was also relaxed because forecast applications based on ARMA models do not require the data to be Gaussian, i.e., ARMA models provide the best linear prediction even for non-Gaussian data (Brockwell and Davis, 1991).

The temporal phase shift exhibited by forecasts obtained by the univariate ARMA model can be partially explained by the error induced by storm movement. One can ameliorate this effect by selecting additional data measured at other neighboring stations (e.g., by using the cross-correlation between the rainfall at the station of

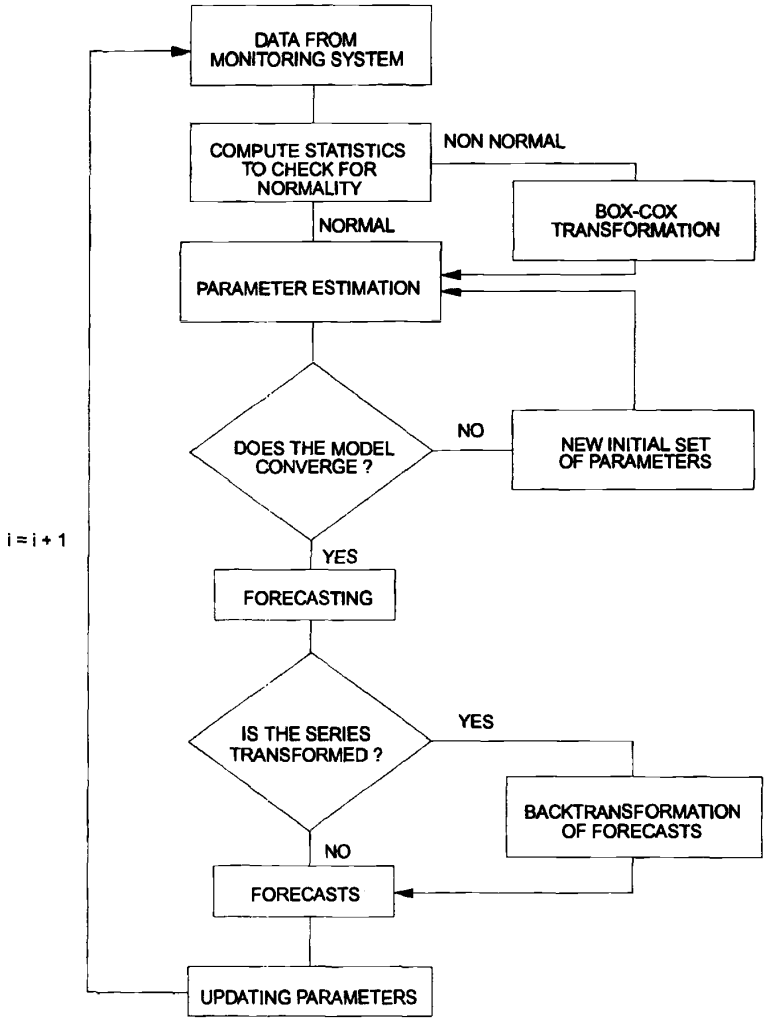


Figure 1 Flowchart of the event-based ARMA forecasting procedure (from Burlando et al., 1993).

interest, i.e., the station where the forecast is issued, and those at the other stations) and reduce the error associated with the phase shift. Using a multivariate integrated ARMA (MARIMA) forecasting scheme (Montanari et al., 1994; Burlando et al., 1996) can do this. Montanari et al. (1994) suggested that a multivariate scheme could remarkably improve the forecasts when the rain gages to be used in forecasting

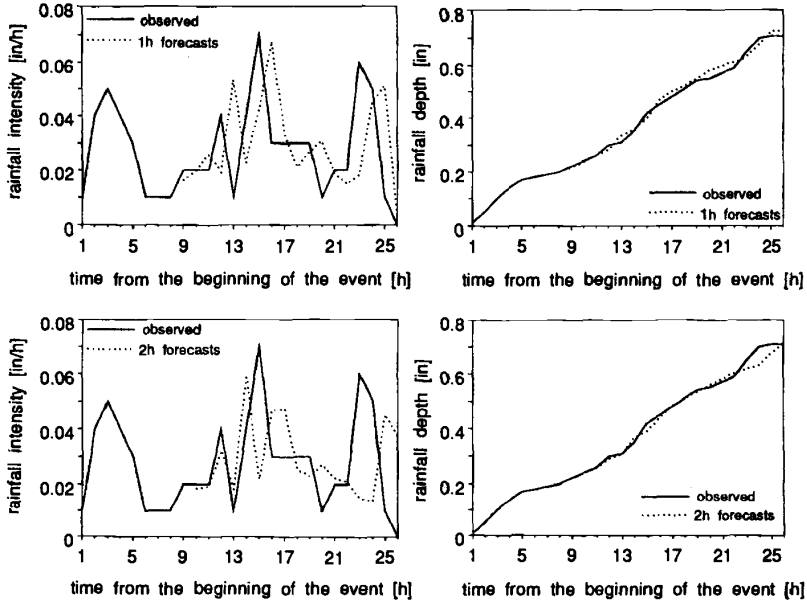


Figure 2 Example of 1- and 2-h rainfall forecasts for the event of October 14, 1960, Denver, Colorado, obtained by means of an ARMA(2,2) process (from Burlando et al., 1993).

are selected adequately. Burlando et al. (1996) showed that the estimation of a Lagrangian space–time correlation of the moving storm could be made using storm maps recorded by weather radar, which provide the direction and the speed of the storm movement. Storm tracking can thus be applied to actual events to select those stations that are characterized by the highest Lagrangian cross-correlation of observed precipitation, and therefore are the best suitable for application with the multivariate model. The parameters of the multivariate model are thus estimated using only observed rainfall at the selected stations throughout the current event.

Specifically, the MARIMA model estimates the future occurrences of a time series as a linear combination of (a) past occurrences of the underlying time series and of time series which are cross-correlated to it—i.e., the autoregressive component—and of (b) the present and past occurrences of a random white-noise component—i.e., the moving average component. The MARIMA model can be expressed as

$$Z_t = \sum_{i=1}^{p} \Phi_i Z_{t-i} + \sum_{j=0}^{q} \Theta_j \varepsilon_{t-j} \tag{6}$$

where p and q are the autoregressive and the moving average order respectively, $\mathbf{Z}_t = (\mathbf{I} - \mathbf{B})^d \mathbf{X}_t$, $\mathbf{X}_t$ is the rainfall intensity, $\mathbf{I}$ is the identity matrix, $\mathbf{B}$ is the backward operator, d is the differencing order of the model, and ε_t is a normally distributed noise term. Both $\mathbf{Z}_t$ and $\mathbf{X}_t$ are n-dimension column vectors (n = number of series), and Φ and Θ are the $n \times n$ autoregressive and moving average parameters matrices of the model. The number of parameters in (6) becomes large as the orders p and q increase. This is a major limitation in analytical tractability and parameter estimation especially in those cases where a limited number of observations are available. Accordingly, the values of p and q, as well as the number of series n, should be selected as a compromise between the conflicting needs of the process descriptiveness and of mathematical tractability.

Burlando et al. (1996) explored the suitability of the MARIMA(1,1,0) model for a catchment in northern Italy. Parameter estimation was carried out on individual events, as in Burlando et al. (1993), and using the method of moments as

$$\Phi = \mathbf{M}_1 \mathbf{M}_0^{-1} \tag{7a}$$

$$\Theta\Theta^T = \mathbf{M}_0 - \mathbf{M}_1 \mathbf{M}_0^{-1} \mathbf{M}_1^T \tag{7b}$$

where $\mathbf{M}_0$ and $\mathbf{M}_1$ denote the lag-0 and lag-1 covariances, respectively. The identification of the pair of stations was carried out either on the basis of historical cross-correlations or from the analysis carried out in real time from radar maps. The latter provided the basis for the analysis of the kinematic characteristics of the storm, so allowing the identification of a (first) *lead station*, located downwind the (second) forecasting *station*. The lead station is taken as a reference station for the second station is selected among those located along the direction of the storm movement that is identified from the radar maps. The MARIMA(1,1,0) was thus estimated using rain gage data observed at the selected stations, and rainfall forecasts were issued at each station as a function of the current and past occurrences observed at the station itself and at the lead station. Satisfactory results were obtained as reported in Burlando et al. (1996).

Artificial Neural Networks. An alternative route to the foregoing stochastic forecasting techniques is the use of artificial neural networks. These are essentially data processing systems that can reproduce by learning the relationships between a pair of one- or multidimensional data sets. An artificial neural network (ANN) is made of many simple nonlinear units that mimic the human neurons. These collect the input from a single or multiple sources producing an output according to a predefined nonlinear function. In a sense an ANN is a sort of a transfer function model that appears to be suitable to tackle the problem of rainfall forecasting.

Use of ANNs for the purpose of weather-related quantities started in the early 1990s. French et al. (1992b) developed a neural network to forecast rainfall intensity fields in time and space, which were generated by a modified version of the stochastic rainfall simulation model proposed by Rodriguez-Iturbe and Eagleson (1987). The network with input, hidden, and output layers was trained using the back-

propagation technique on a regular grid domain to test the ability of the ANN to investigate the role of the number of hidden nodes on its performance. The model skill was tested based on a varying number of training sets and the rainfall fields generated by the stochastic model. Real-time learning and off-line learning were additionally tested. Kuligowksi and Barros (1998) applied a combination of precipitation data from a number of rain gages and wind directions to forecast rainfall amounts for a target location and a lead time of 6 h. Specifically, rainfall observations at rain gages in a region of radius 300 km centered on the target location, upper level winds at three radiosonde locations, and wind direction data from a number of levels were combined to build the training set of the ANN.

More recently, Luk et al. (2000) adopted ANNs to forecast short-term rainfall for an urban catchment, aiming at the investigation of the effect of temporal and spatial information on short-term rainfall forecasting. The forecast accuracy of ANNs was evaluated for different configurations of lag orders and number of spatial inputs based on historical rainfall patterns. They concluded that the most accurate predictions depend on the identification of an optimum number of spatial inputs, and that the network with lower lag consistently produced better performance. An interesting application of ANNs has been recently shown by Toth et al. (2000), who provided for a real case study a comparison of ANNs performance with respect to real-time prediction based on ARMA models and a nonparametric nearest-neighbor technique. Multilayer feed-forward network architectures were tested against the one-layer scheme in order to determine the optimal network configuration, both in the case of split-sample application and adaptive calibration. As one may expect, better performances were obtained for the split-sample application, which makes use of larger training sets, whereas the adaptive calibration gives worse results for short lead times. Compared to other forecasting techniques ANNs was slightly superior in the overall performance due to their ability to account for the nonlinearities that characterize temporal rainfall. Grecu and Krajewski (2000) reported another interesting application of back-propagation neural network (BPNN) for rainfall forecasting. In this case, rainfall amounts were not directly modeled by means of the BPNN, but this was used to model one component of the statistical radar-based quantitative precipitation forecast procedure.

Mid- and Long-Term Forecasting

If quantitative short-term forecasting is useful for flood forecasting, mid- and long-term forecasting plays a major role in the management of water resources. Agriculture and water supply, among other water uses, can significantly benefit from the availability of forecasts of rainfall amounts that can be expected over a time horizon of a month or a season. This issue is particularly relevant for complex systems that strongly depend on joint management of surface and groundwater resources. Forecasting at the mid- and long-term scales involves problems that are similar to the one already observed for smaller time scales. Nonstationarity, nonlinearity as well as the identification of the correct predictors guided the development of methods.

Whereas regression methods and, more recently, artificial neural networks have been extensively used for the purpose, a few other approaches can be found in the

literature to forecast rainfall on mid- and long-term time scales. A truncated normal distribution is, for instance, the basis of the formulation of a nonstationary multisite model of rainfall that Sansó and Guenni (2000) show to capture the year-to-year variability and suggest to be suitable for short-term forecasting as well. Stone et al. (1996) and de Jager et al. (1998) used a simple probabilistic rainfall forecasting technique that is based on the identification of lag relationships between the values of the Southern Oscillation Index (SOI)—which can be considered as representative of the phase of the El Niño Southern Oscillation (ENSO) cycle—and future rainfall. Probability distributions for the subsequent 3 months are thus derived conditioned on the state of the SOI. Sharma (2000) introduced a nonparametric probabilistic model for forecasting rainfall with 3 to 24 months of lead times. Specifically, nonparametric kernel methods (e.g., Scott, 1992) for probability density function (PDF) estimation are used to express the conditional probability density function. Then, probabilistic forecasts are made by resampling from the rainfall probability density conditioned on the current value of the associated predictor set. An interesting feature of this approach is that the shape of the PDF is directly built from the data, and this leads to forecasts that are expected to resemble the characteristics of the sample and therefore reproduce the variability of observed rainfall.

Regression-based techniques have been extensively used for predicting seasonal rainfall. The increased availability of predictor variables, like ENSO, in near real time, by means of either observations or numerical weather prediction has increased the applicability of regression-based models and the Kalman filter (e.g., Liu et al., 1998). Makarau and Jury (1997) forecasted summer rainfall in Zimbabwe based on a set of climatic predictors by means of a multivariate linear regression model in a forward stepwise approach. Fairly simple models including up to five predictors produced jack-knife skill test correlations of about 80 to 85% for a lead time of 2 to 3 months. A similar approach was used by Jury (1998) to forecast seasonal rainfall and other climatic variables for the KwalaZulu-Natal region in southern Africa, also obtaining a forecast skill of about 76% for rainfall and about two thirds of the variance in the other cases. Francis and Renwick (1998) focused on predicting seasonal (either 1 month or one 3-month season) rainfall anomalies.

Similarly, Thapliyal (1997) carried out a comparison of forecast models based on the correlation between predictors and predictands and a dynamic stochastic transfer model to predict monsoon rainfall in India. The dynamic stochastic transfer model corresponds essentially to an ARIMA model structure, the orders of which are estimated against observations. It should be observed that a critical issue in using regression-based techniques is the stability of the selected predictor and the robustness of the model describing its temporal evolution. Finally, ARMA models, which have been extensively applied to forecast streamflows, have also been used to simulate rather than to forecast mid- and long-term rainfall. Another application of ARIMA models for forecasting monthly rainfall series with the purpose of providing an input for flow forecasting in the management of water resources systems can be found in Delleur and Kavvas (1978).

As in the case of short-term forecasts, artificial neural networks have been proposed for forecasting seasonal rainfall. ANNs has been found to be useful for forecasting the behavior of complex and highly dynamic systems such as the

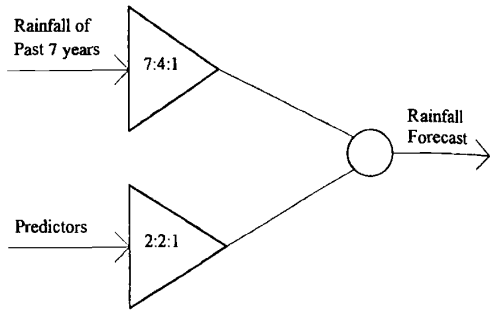


Figure 3 Architecture of a hierarchical artificial neural network combining deterministic information from historical data and stochastic component represented by the event predictors for seasonal forecasting of monsoon rainfall (from Navone and Ceccatto, 1994).

monsoon rainfall. Simple deterministic neural networks show, however, a limited robustness in terms of forecast skill, so that more complex networks are often introduced. An example of a simple four-layer input, two hidden layers, and one neuron output network for forecasting Indian monsoon rainfall can be found in Sahai et al. (2000). On the other hand, Navone and Ceccatto (1994), proposed an interesting but complex application of ANNs, as a nonlinear method to correlate preseason predictors to rainfall data and as an algorithm for reconstructing the rainfall time series dynamics. Accordingly, they implemented a hierarchical neural network, which is sketched in Figure 3. The network trained to correlate predictors and the network trained to learn the time series dynamics are combined by connecting their output units to a new neuron, which is then used to issue the forecasts. The authors refer about an improved forecast skill due to the hierarchical approach, especially in forecasting large anomalies. The performance of ANNs can be reduced if the parameters used to train and forecast are correlated. Guhathakurta et al. (1999) used principal component analysis—as suggested by Hsieh and Tang (1998)—to transform the original variables into a new set of uncorrelated variables. These are then used to train and issue forecasts by means of a three-layer, five-input, three hidden nodes in one single hidden layer ANN. The output from such ANN and the output of a simple deterministic ANN using the untransformed parameter set were then used each as input to a simple two-layer ANN without any hidden layer, which produced rainfall forecasts. The final hybrid model increased the overall forecast skill from about 40 to 80%.

4 STOCHASTIC STREAMFLOW FORECASTING

Over the past two decades, considerable research has been carried out in hydrology for developing stochastic models for short- and long-term forecasting of river flows.

The form of these models generally follows the ARMA, ARMAX, and transfer function type of models (Box and Jenkins, 1976), with the last two proving to be more reliable for multiple forecasting periods (Burns and McBean, 1985; Awwad and Valdés, 1992). After defining the mathematical model, usually in a state-space form, the Kalman filter is widely used as a powerful tool for obtaining optimal hydrologic forecasts and updates of the states (e.g., Chiu, 1978; O'Connell, 1980; Wood and O'Connell, 1985).

Implementation of real-time stochastic models in large-scale hydrologic systems has been thoroughly discussed by Wood (1985) and Awwad and Valdés (1992). Furthermore, adaptive filters have also been used in combination with conceptual hydrologic models for streamflow forecasting since the late 1970s. Bras and collaborators at MIT, developed techniques to represent the National Weather Service River Forecasting System (NWSRFS) land component in a state-space form. Some of the research results are discussed in the next section. Alternative techniques such as stepwise regression and transfer function models have also been used. Later in this chapter the application of ANN in streamflow forecasting will be presented. In addition, probabilistic, interval, forecasts using the standard conceptual rainfall–runoff models for short-term forecasting has been developed. The most widely used is the ESP (extended streamflow prediction, now called ensemble streamflow prediction), which is described later in this chapter.

Short-Term Forecasting of Streamflows

A number of investigators have evaluated the benefits of streamflow forecast using the Kalman filter. For example, Georgakakos (1986) studied the performance of a hydrometeorological model for streamflow forecasting using 6-h data for Bird Creek, a 2344 km^2 catchment in Oklahoma. A precipitation forecasting model was developed and coupled to a modified version of the U.S. National Weather Service rainfall–runoff model to produce the streamflow forecast. Variables such as soil moisture storages and streamflow were updated through a Kalman filter. Eight 2-month forecast periods were examined. The results showed that the nonupdating forecasting model produced forecasts where the time-to-peak discharges were very different from those observed. The forecasts using updating techniques showed significant improvements.

Other applications in streamflow forecasting include the work of Takasao and Shiiba (1984) and Takasao et al. (1989) who developed a simple nonlinear streamflow forecasting model and applied to the Haze River, a 370-km^2 basin in Japan. Their work shows model performance for a flood in September 1965 with and without updating. As expected, the forecast errors of the model without updating are larger. The deterministic model NAMS11/MIKE11 developed by the Danish Hydraulic Institute (DHI) uses a state variable updating procedure based on the Kalman filter for their conceptual rainfall–runoff model NAM (Refsgaard, 1997). Ahsan and O'Connor (1994) have expressed that the full capabilities of the Kalman filtering are not completely utilized since the predictions are expected to match the

observed flows, which are considered to be noise free and that the filter will be more fully utilized in the future when remote sensing becomes more predominant.

Awwad and Valdés (1992) proposed an adaptive evaluation/forecasting algorithm for hydrologic forecasting and presented two multisite hydrologic forecasting approaches suitable for real-time applications. Their model is based on past and present flow rates, with the upstream inflows treated as exogenous inputs to the models. They applied the model to the Fraser River, Canada. In their original application Awwad and Valdés (1992) did not use precipitation terms, and even though their models performed very well in the one- and two-steps-ahead forecast error deteriorated rapidly. The authors later extended the adaptive evaluation/forecasting algorithm to include precipitation inputs, upstream inflows forecasted/evaluated with uncertainty, and deterministic reservoir releases in the stochastic models (Awwad et al., 1994). This approach was adopted because it has a relatively simple dynamic structure in a black-box form and is calibrated online as additional information becomes available. The inclusion of precipitation information considerably benefited the multiple-period forecasting ability of the stochastic models.

The general form of the ARMAX models used in Awwad et al. (1994) followed the well-known state-space form of the Kalman filter presented above where optimal forecasts and updates of the states were obtained using the Kalman filter. Two other filters in the form of the Kalman filter, referred to as the parameter space and the noise-space filters, are used in parallel with the state-space filter to update the model parameters and noise statistics online along with the states. This adaptive estimation technique using parallel filtering does not require preassigned values for the Kalman filter coefficients and noise statistics, which are usually unknown in real-world applications. Other applications in short-term forecasting include: use of an ARMAX model to do predictions on the Fraser River in Canada (Ngan and Russell, 1986), use of the Kalman filter to estimate the parameters of a PARMA model with application to the Saugeen River in Ontario, Canada (Jimenez et al., 1989), and use of the ARMAX model with Kalman filter for short-term flow forecasting of snowmelt runoff in the Rio Grande at the Del Norte station (Haltiner and Salas, 1988).

As stated before in section 3 ANNs have been widely used for a number of hydrologic problems including forecasting of precipitation and streamflow. A vast literature already exists on the subject. For example, the *ASCE J. Hydrol. Engr.* (vol. 5, no. 2, 2000) is a dedicated issue on the subject. It includes the articles "Artificial Neural Networks in Hydrology I: Preliminary Concepts" and "Artificial Neural Networks in Hydrology II: Hydrologic Application," co-authored by the ASCE Task Committee on Applications of Artificial Neural Networks in Hydrology. The second article includes a review of various applications of ANNs on short-term and long-term flow forecasting. In addition, the book *Artificial Neural Networks in Hydrology* (Govindaraju and Rao, 2000) includes some chapters specifically on flow forecasting (e.g., Gupta et al., 2000; Salas et al., 2000; Deo and Thirumalaiah, 2000). The work by Gupta et al. (2000) discusses in some detail the training of ANNs based on multilayer feedforward neural networks (MFNNs), which are most commonly used for streamflow forecasting. It also presents some results illustrating

some applications. The study by Salas et al. (2000) discuss some very basic concepts underlying ANNs, gives a simple detailed example, and two applications for daily and monthly streamflow forecasting. Finally, the work by Deo and Thirumalaiah (2000) includes the application of ANNs for real-time forecasting of daily flows and daily river stages.

Long-Term Forecasting of Streamflows

Forecasting Models Using Climatic Precursors. Long-term forecasting of hydrologic variables requires climate forecasts. As mentioned in the section on long-term precipitation forecasting, the increase in predictive skills of the models to forecast climatic anomalies based on the ENSO phenomena have provided renewed impetus for hydrologic forecasting.

There have been a significant number of contributions to hydrologic forecasting using climatic precursors. For example, the hydrologic forecasting model proposed by Liu et al. (1997, 1998) used multiple ENSO forecasts to produce forecasts of seasonal precipitation, streamflow, and other variables. This is essentially a combined data fusion and forecasting system that incorporates ENSO forecasts, persistence-based forecasts, and up-to-date observations. The system assimilates past observations, hindcasts, and projects with both multiple model outputs and persistence into its merged forecast. The error bounds on the forecast are also propagated. Liu and co-workers applied the approach to forecast droughts in Texas (Liu et al., 1997) and precipitation in the Equatorial Pacific and streamflows in Colombia (Liu et al., 1998). An example of these applications is given in Figure 4. Other applications include Berri and Flamenco (1999) who used a regression model with climatic precursors to forecast seasonal volumes in the Rio Diamante and Chiew et al. (1998) who studied the relationship between ENSO and rainfall, drought and flows in Australia, and its potential for forecasting. For example, they found that spring runoff in southeast Australia may be predicted several months in advance. Other examples are the work of Guetter and Georgakakos (1996) on the relationship between Iowa River flows and ENSO and of Kayha and Dracup (1993, 1994) on the Southwest flows.

An interesting example of the use of seasonal climate outlooks of expected air temperature and precipitation probabilities in hydrologic forecasting is found in the work of Croley and colleagues at the NOAA GREL (Croley, 1996, 1997; Croley and Kunkel, 1996). In their work historical meteorology record segments are used with hydrological and other models to simulate hydrological scenarios. The historical meteorological records are weighted to be compatible with NOAA's climate outlooks. An example of the use of the procedure is shown in Figure 5. Hamlet and Lettenmaier (1999) proposed a method to incorporate both ENSO and PDO signals in the long-term forecasting of streamflows in the Columbia River. The climate forecasts, classified in six categories as a function of ENSO and PDO, are used to generate climatic scenarios that serve as input to a hydrologic model.

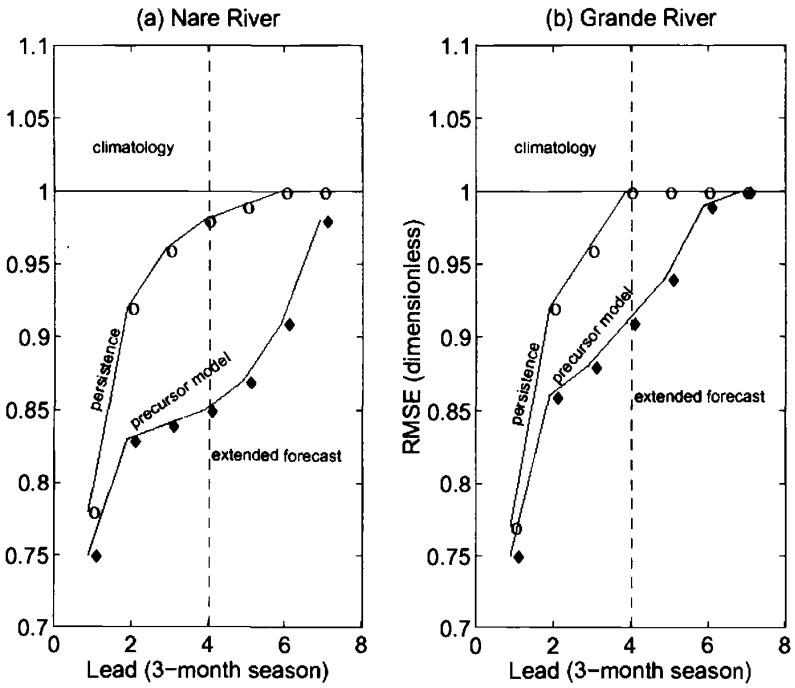


Figure 4 Performance of long-term streamflow forecasting models using climatic precursors (from Liu et al., 1998).

Extended/Ensemble Streamflow Prediction (ESP). The extended streamflow prediction (ESP) was proposed by the National Weather Service (Twedt et al., 1977; Day, 1985) as a procedure to obtain probable runoff scenarios based on simulations runs of a conceptual simulation model that uses hydroclimatic series, either historic or synthetic, and using the current soil conditions. The hydroclimatic series, usually temperature and precipitation, represent historic or synthetic time series that are likely to happen in the period of forecast. The results are then used to obtain a probability distribution of future runoff and, if desired, a point estimate, usually the sample average for a particular lead time. Figure 6 shows an example for the Rex River in Washington (Lettenmaier and Wood, 1993). The ESP approach has been widely used in the United States and other countries as a component of their forecast systems, and it is usually combined with a reservoir operation model that uses the probabilistic outputs of the ESP approach. For example, as part of the NWSRFS the ESP approach is being utilized in the operation of the Columbia River (260,000 mi^2 and a mean annual discharge of 198 million acre-feet; von der Heydt et al., 1994).

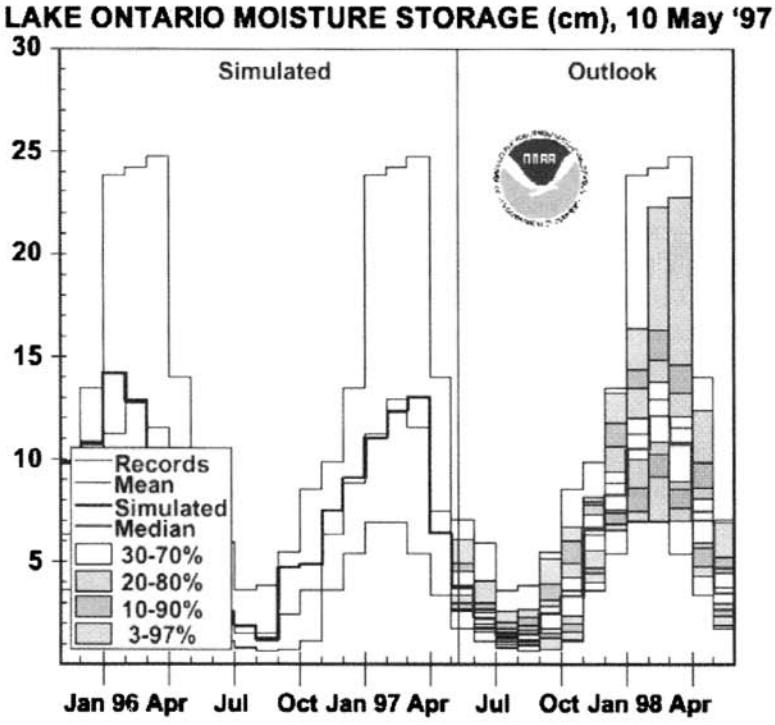


Figure 5 Example of hydrologic forecasts using climate outlooks (from Croley, 1997, 2000). See ftp site for color image.

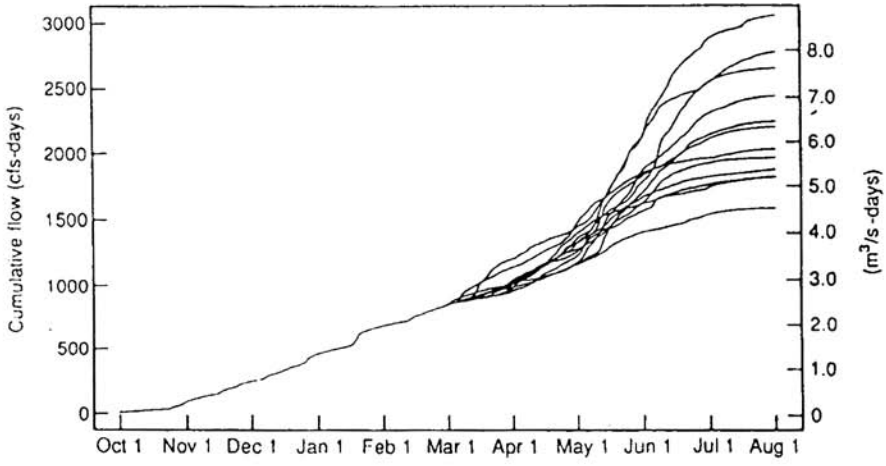


Figure 6 Example of ensemble streamflow prediction application (taken from Lettenmaier and Wood, 1993).

REFERENCES

Ahsan, M., and K. M. O'Connor, A simple non-linear rainfall-runoff model with a variable gain factor, *J. Hydrol.*, *155*, 151–183, 1994.

Antolik, M. S., An overview of the National Weather Service's Centralized Statistical Quantitative Precipitation Forecasts, *J. Hydrol.*, *239*, 306–337, 2000.

Awwad, H. M., and J. B. Valdes, Adaptive parameter-estimation for multisite hydrologic forecasting, *J. Hydraul. Eng. ASCE*, *118*(9), 1201–1221, 1992.

Awwad, H. M., J. B. Valdes, and P. J. Restrepo, Streamflow forecasting for Han River Basin, Korea, *ASCE J. Water Resour. Pl.*, *120*(5), 651–673, 1994.

Berri, G. J., and E. Flamenco, Seasonal volume forecast in the Diamante River, Argentina based on El Niño observations and predictions, *Water Resour. Res.*, *35*(12), 3803–3810, 1999.

Bertoni, J. C., C. E. Tucci, and R. T. Clarke, Rainfall-based real-time flood forecasting, *J. Hydrol.*, *131*, 313–339, 1992.

Bougeault, P., P. Binder, A. Buzzi, R. Dirks, R. Houze, J. Kuettner, R. B. Smith, R. Steinacker, and H. Volkert, The MAP special observing period, *Bull. Am. Met. Soc.*, *82*(3), 433–462, 2000.

Box, G. E. P., and D. R. Cox, An analysis of transformations, *J. R. Statist. Soc. B*, *26*, 211–252, 1964.

Box, G. E. P., and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, Holden-Day Press, San Francisco, 1976.

Brath, A., P. Burlando, and R. Rosso, Sensitivity analysis of real-time flood forecasting to on-line rainfall predictions, in F. Siccadi and R. L. Bras (Eds.), *Selected Papers of the Workshop on Natural Disasters in European Mediterranean Countries*, Colombella, Perugia, Italy, 1988, pp. 469–488.

Brockwell, P. J., and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed., Springer-Verlag, New York, 1991.

Brown, G. R., and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, 3rd ed., Wiley, New York, 1996.

Burlando, P., and R. Rosso, Stochastic models of temporal rainfall: Reproducibility, estimation and prediction of extreme events, in J. Marco-Segura, R. Harboe, and J. D. Salas (Eds.), *Stochastic Hydrology and Its Use in Water Resources Systems Simulation and Optimization*, Kluwer, Dordrecht, 1993, pp. 137–173.

Burlando, P., R. Rosso, L. Cadavid, and J. D. Salas, Forecasting of short-term rainfall using ARMA models, *J. Hydrol.*, *144*, 193–211, 1993.

Burlando, P., A. Montanari, and R. Ranzi, Forecasting of storm rainfall by combined use of radar, rain gages and linear models, *Atmos. Res.*, *42*, 199–216, 1996.

Burn, D. H., and E. A. McBean, River flow forecasting model for Sturgeon river, *J. Hydraul. Eng. ASCE*, *111*(2), 316–333, 1985.

Carter, M. M., and J. B. Elsner, A statistical method for forecasting rainfall over Puerto Rico, weather and forecasting, *12*(3), 515–525, 1997.

Carter, M. M., J. B. Elsner, and S. P. Bennett, A quantitative precipitation forecast experiment for Puerto Rico, *J. Hydrol.*, *239*, 162–178, 2000.

Chiew, F. H. S., T. C. Piechota, J. A. Dracup, and T. A. McMahon, El Niño/Southern Oscillation and Australian rainfall, streamflow and drought: Links and potential for forecasting, *J. Hydrol., 204*, 138–149, 1998.

Chiu, C. L. (Ed.), *Applications of Kalman Filter to Hydrology, Hydraulics and Water Resources*, University of Pittsburgh Press, Pittsburgh, PA, 1978.

Croley II, T. E., *Using Meteorology Probability Forecasts in Operational Hydrology*, American Society of Civil Engineer (ASCE) Press, 2000.

Croley II, T. E., Using NOAA's new climate outlooks in operational hydrology, *ASCE J. Hydrol. Eng., 1*(3), 93–102, 1996.

Croley II, T. E., Water resource predictions from meteorological probability forecasts, in D. Rosbjerg et al., *Proceedings of the Sustainability of Water Resources Under Increasing Uncertainty*, IAHS Publication 240, IAHS Press, Institute of Hydrology, Wallingford, Oxfordshire, 1997, pp. 301–309.

Croley II, T. E., and K. Kunkel, Application of the new NWS climate outlook in operational hydrology, in *Proceedings of the Thirteenth Conference on Probability and Statistics in Atmospheric Sciences*, American Meteorological Society, San Francisco, CA, 1996, pp. 231–238.

Dahale, S. D., and P. V. Puranik, Climatology and predictability of the spatial coverage of 5-day rainfall over Indian subdivisions, *Int. J. Climatol., 20*(4), 443–453, 2000.

Day, G., Extended streamflow forecasting using NWSRFS, *ASCE J. Water Res. Planning Mgmt., 111*(2), 157–170, 1985.

de Jager, J. M., A. B. Potgieter, and W. J. van den Berg, Framework for forecasting the extent and severity of drought in maize in the Free State Province of South Africa, *Agric. Syst., 57*(3), 351–365, 1998.

Delleur, J. W., and M. L. Kavvas, Stochastic models for monthly rainfall forecasting and synthetic generation, *J. Appl. Meteorol., 17*, 1528–1536, 1978.

Deo, M. C., and K. Thirumalaiah, Real time forecasting using neural networks, in R. S. Govindaraju and A. R. Rao (Eds.), *Artificial Neural Networks in Hydrology*, Kluwer Academic, Dordrecht, 2000, pp. 53–72.

Entekhabi, D., I. Rodriguez-Iturbe, and P. S. Eagleson, Probabilistic representation of the temporal rainfall process by a modified Neyman-Scott rectangular pulse model: Parameter estimation validation, *Water Resour. Res., 25*(2), 295–302, 1989.

Fraedrich, K., and K. Müller, On single station forecasting: Sunshine and rainfall Markov chains, *Beitr. Phys. Atmos., 56*, 108–134, 1983.

Francis, R. I. C. C., and J. A. Renwick, A regression-based assessment of the predictability of New Zealand climate anomalies, *Theor. Appl. Climatol., 60*, 21–36, 1998.

French, M. N., R. L. Bras, and W. F. Krajewski, A Monte-Carlo study of rainfall forecasting with a stochastic model, *Stochast. Hydrol. Hydraul., 6*(1), 27–45, 1992a.

French, M. N., W. F. Krajewski, and R. R. Cuykendall, Rainfall forecasting in space and time using a neural network, *J. Hydrol., 137*, 1–31, 1992b.

French, M. N., and W. F. Krajewski, A model for real-time quantitative rainfall forecasting using remote sensing. 1. Formulation, *Water Resour. Res., 30*(4), 1075–1083, 1994.

French, M. N., H. Andrieu, and W. F. Krajewski, A model for real-time quantitative rainfall forecasting using remote sensing. 1. Formulation, *Water Resour. Res., 30*(4), 1085–1097, 1994.

Georgakakos, K. P., A generalized stochastic hydrometeorological model for flood and flash-flood forecasting: 2. Case studies, *Water Resour. Res.*, *22*(13), 2096–2106, 1986.

Glahn, H. R., and D. A. Lowry, The use of model output statistics (MOS) in objective weather forecasting, *J. Appl. Meteorol.*, *11*, 1203–1211, 1972.

Govindaraju, R. S., and A. R. Rao, *Artificial Neural Networks in Hydrology*, Kluwer Academic, Dordrecht, 2000.

Gray, W. M., W. L. Christofer, P. W. Mielke, and K. R. J. Berry, Predicting Atlantic Basin seasonal tropical cyclone activity by 1 June, *Weather Forecast.*, *9*, 103–115, 1994.

Grecu, M., and W. Krajewski, Simulation study of the effects of model uncertainty in variational assimilation of radar data on rainfall forecasting, *J. Hydrol.*, *239*(1–4), 85–96, 2000.

Gregory, J. M., T. M. Wigley, and P. D. Jones, Determining and interpreting the order of a two-state Markov chain: Application to models of daily precipitation, *Water Resour. Res.*, *28*(5), 1443–1446, 1992.

Gupta, H. V., K. Hsu, and S. Sorooshian, Effective and efficient modeling for streamflow forecasting, in R. S. Govindaraju, and A. R. Rao (Eds.), *Artificial Neural Networks in Hydrology*, Kluwer Academic, Dordrecht, 2000, pp. 7–22.

Guetter, A. K., and K. P. Georgakakos, Are the El Niño and La Niña predictors of the Iowa River seasonal flow, *J. Appl. Meteorol.*, *35*, 690–705, 1996.

Guhathakurta, P., M. Rajeevan, and V. Thapliyal, Long range forecasting Indian summer monsoon rainfall by a hybrid principal component neural network model, *Meteorol. Atmos. Phys.*, *71*(3–4), 255–266, 1999.

Halinter, J. P., and J. D. Salas, Short-term forecasting of snowmelt runoff using ARMAX models, *Water Resour. Bull.*, *24*(5), 1083–1089, 1988.

Hamlet, A. F., and D. P. Lettenmaier, Columbia River streamflow forecasting based on ENSO and PDO climate signals, *ASCE J. Water Resour. Planning Mgmt.*, *125*(6), 333–341, 1999.

Hsieh, W. W., and B. Tang, Applying neural network models to prediction and data analysis in meteorology and oceanography, *Bull. Am. Meteorol. Soc.*, *79*, 1855–1870, 1998.

Jimenez, C., A. I. McLeod, and K. W. Hipel, Kalman filter estimation for periodic auto-regressive-moving average models, *Stochastic Hydrol. Hydraul.*, 227–240, 1989.

Jinno, K., A. Kawamura, R. Berndtsson, M. Larson, and J. Niemczynowicz, Real-time rainfall prediction at small space-time scales using a 2-dimensional stochastic advection-diffusion model, *Water Resour. Res.*, *29*(5), 1489–1504, 1993.

Johnson, E. R., and R. L. Bras, Multivariate short-term rainfall prediction, *Water Resour. Res.*, *16*(1), 173–185, 1980.

Jury, M. R., Statistical analysis and prediction of KwaZulu-Natal climate, *Theor. Appl. Climatol.*, *60*(1–4), 1–10, 1998.

Kalman, R. E., A new approach to linear filtering and prediction problems, *Trans. ASME J. Basic Eng. Ser. D*, *82*, 35–45, 1960.

Kalman, R. E., and R. S. Bucy, New results in linear filtering and prediction theory, *Trans. ASME J. Basic Eng. Ser. D*, *83*, 95–107, 1961.

Kayha, E., and J. A. Dracup, US streamflow patterns in relation to the El Niño Southern Oscillation, *Water Resour. Res.*, *29*, 2491–2503, 1993.

Kayha, E., and J. A. Dracup, The influences of type 1 El Niño and La Niña events on streamflows in the Pacific Southwest of the United States, *J. Climatol.*, *7*, 965–976, 1994.

Katz, R. W., On some criteria for estimating the order of a Markov chain, *Technometrics*, *23*(3), 243–249, 1981.

Kawamura, A., K. Jinno, R. Berndtsson, and T. Furukawa, Parameterization of rain cell properties using an advection-diffusion model and rain gage data, *Atmos. Res.*, *42*, 67–73, 1996.

Kawamura, A., K. Jinno, R. Berndtsson, and T. Furukawa, Real-time tracking of convective rainfall properties using a two-dimensional advection-diffusion model, *J. Hydrol.*, *203*(1–4), 109–118, 1997.

Kuligowski, R. J., and A. P. Barros, Experiments in short-term precipitation forecasting using artificial neural networks, *Monthly Weather Rev.*, *126*, 470–482, 1998.

Labadie, J. W., R. C. Lazaro, and D. M. Morrow, Worth of short-term rainfall forecasting for combined sewer overflow control, *Water Resour. Res.*, *17*(5), 1489–1497, 1981.

Lardet, P., and C. Obled, Real-time flood forecasting using a stochastic rainfall generator, *J. Hydrol.*, *162*(3–4), 391–408, 1994.

Lettenmaier, D. P., and E. F. Wood, Hydrologic forecasting, in D. R. Maidment (Ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993.

Liu, Z., J. B. Valdés, and D. Entekhabi, Merged forecasts of drought index anomalies along the Gulf Coast in the US using multiple precursors, *Exper. Long-Lead Forecast Bull.*, *6*(2), 9–11, 1997.

Liu, Z., J. B. Valdés, and D. Entekhabi, Merging and error analysis of regional hydrometeorologic anomaly forecasts conditioned on climate precursors, *Water Resour. Res.*, *34*(8), 1959–1969, 1998.

Lowry, D. A., and H. R. Glahn, An operational model for forecasting probability of precipitation—PEATMOS PoP, *Monthly Weather Rev.*, *104*, 221–232, 1976.

Luk, K. C., J. E. Ball, and A. Sharma, A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting, *J. Hydrol.*, *227*(1–4), 56–65, 2000.

Makarau, A., and M. R. Jury, Predictability of Zimbabwe summer rainfall, *Int. J. Climatol.*, *17*(13), 1421–1432, 1997.

Miller, A. J., and L. M. Leslie, Short-term single station forecasting of precipitation, *Monthly Weather Rev.*, *112*, 1198–1205, 1984.

Montanari, A., P. Burlando, and R. Rosso, Forecasting of short-term rainfall using multivariate ARMA models, *Annal. Geophys.*, *12*(sp. issue), C325–C409, 1994 (abstract).

Navone, H. D., and H. A. Ceccatto, Predicting Indian monsoon rainfall—a neural network approach, *Climate Dynam.*, *10*(6–7), 305–312, 1994.

Ngan, P., and S. O. Russell, Example of flow forecasting with Kalman filter, *ASCE J. Hydraul. Eng.*, *112*(9), 818–832, 1986.

Nguyen, V. T. V., M. B. McPherson, and J. Rousselle, *Urban Water Resource Research Program*, Technical Memo 35, American Society of Chemical Engineers, New York, 1978.

Obeysekera, J. T. B., G. Q. Tabios III, and J. D. Salas, On parameter estimation of temporal rainfall models, *Water Resour. Res.*, *23*(10), 1837–1850, 1987.

O'Connell, P. E. (Ed.), *Real Time Hydrological Forecasting and Control*, Institute of Hydrology, Wallingford, England, 1980.

Phanartzis, C. A., Rainfall prediction, Progress Report Wastewater Program, City and County of San Francisco, CA, 1979.

Ramirez, J. A., and R. L. Bras, Conditional distributions of Neyman-Scott models for storm arrivals and their use in irrigation control, *Water Resour. Res.*, *21*, 317–330, 1985.

Refsgaard, J. C., Validation and intercomparison of different updating procedures for real-time forecasting, *Nordic Hydrol.*, *28*, 65–84, 1997.

Rodriguez-Iturbe, I., and P. S. Eagleson, Mathematical models of rainstorm events in space and time, *Water Resour. Res.*, *23*(1), 181–190, 1987.

Sahai, A. K., M. K. Soman, and V. Satyan, All India summer monsoon rainfall prediction using an artificial neural network, *Climate Dynam.*, *16*(4), 291–302, 2000.

Salas, J. D., M. Markus, and A. S. Tokar, Streamflow forecasting based on artificial neural networks, in R. S. Govindaraju, and A. R. Rao (Eds.), *Artificial Neural Networks in Hydrology*, Kluwer Academic, Dordrecht, 2000, pp. 23–52.

Sansó, B., and L. Guenni, A nonstationary multisite model for rainfall, *J. Am. Statist. Assoc.*, *95*(452), 1089–1100, 2000.

Schaake, J. C., and L. Larson, A strategy for ensemble streamflow prediction (ESP), *Proceedings of the American Meteorological Society Annual Meeting*, Phoenix, AZ, Vol. J104-J105, 1997.

Scott, D. W., *Multivariate density estimation: Theory, practice and visualisation*. Probability and Mathematical Statistics Series, Wiley, New York, 1992.

Sharma, A., Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3—a non parametric probabilistic forecast model, *J. Hydrol.*, *239*(1–4), 249–258, 2000.

Stengel, R. F., *Stochastic Optimal Control Theory and Application*, Wiley, New York, 1986.

Stone, R. C., G. L. Hammer, and T. Marcussen, Prediction of global rainfall using phases of the Southern Oscillation index, *Nature*, *384*, 252–255, 1996.

Sugimoto, S., E. Nakakita, and S. Ikebuchi, A stochastic approach to short-term rainfall prediction using a physically based conceptual rainfall model, *J. Hydrol.*, *242*(1–2), 137–155, 2001.

Takasao, T., and M. Shiiba, Development of techniques for on-line forecasting of rainfall and flood runoff, *Natural Disaster Sci.*, *6*(2), 83–112, 1984.

Takasao, T., M. Shiiba, and K. Takara, Stochastic state-space techniques for flood runoff forecasting, in *Proc. Pacific Int. Seminar on Water Resour. Systems*, Tomanu, Japan, 1989, 117–132.

Thapliyal, V., Preliminary and final long range forecast for seasonal monsoon rainfall over India, *J. Arid Environ.*, *36*(3), 385–403, 1997.

Tong, H., Determination of the order of a Markov chain by Akaike's information criterion, *J. Appl. Prob.*, *12*, 488–497, 1975.

Toth, E., A. Brath, and A. Montanari, Comparison of short-term rainfall prediction models for real-time flood forecasting, *ASCE J. Hydrol.*, *239*(1–4), 132–147, 2000.

Trotta, P. D., J. W. Labadie, and N. S. Grigg, Automatic control strategies for urban stormwater, *ASCE J. Hydraul. Div.*, *103*(HY12), 1977.

Twedt, T. M., J. C. Schaake, Jr., and E. L. Peck, National Weather Service extended streamflow prediction, *Proc. 45th Western Snow Conference*, Albuquerque, NM, April 1977, pp. 52–57.

von der Heydt, L., L. E. Brazil, and K. Jawed, ABPA realtime hydrologic forecasting for the Columbia River Basin, in *Proceedings of the 21st Annual Conference of the ASCE Water Resources Planning and Management Division*, Denver, CO, 1994.

Wood, E.F. and P.E. O'Connell, "Real Time Forecasting" in *Hydrological Forecasting*, M. G. Anderson and T. P. Burt (Eds), John Wiley & Sons Inc., 1985, 505–558.

Yu, P. S., and T. C. Yang, A probability-based renewal rainfall model for flow forecasting, *Nat. Hazards*, *15*(1), 51–70, 1997.

Zawadzki, I., Fractal structure and exponential decorrelation in rain, *J. Geophys. Res.*, *92*(D8), 9586–9590, 1987.

CHAPTER 35

REMOTE SENSING AND GEOGRAPHICAL INFORMATION SYSTEMS APPLICATIONS IN HYDROLOGY

EDWIN T. ENGMAN AND NANDISH MATTIKALLI

1 INTRODUCTION

Remote sensing and associated image-processing technology provide access to spatial and temporal hydrologic information from watershed to global scales. Advances in sensor and imaging technology are increasing the capability of remote sensing for specific hydrologic application.

There are two general areas where remote sensing can be used in hydrologic modeling: (1) determining watershed geometry, drainage network, and other map-type information for distributed hydrologic models and for empirical flood peak, annual runoff, or low-flow equations; and (2) providing input data such as snow cover or precipitation, diagnostic variables such as soil moisture or surface temperature, or model parameters such as delineated land-use classes used to define runoff coefficients. In this review, the latter is addressed. The various uses of remote sensing to provide input data and diagnostic variables for hydrologic models are treated as they are used to measure the different hydrologic variables or processes, e.g., precipitation, snow, or evaporation. Each of these hydrologic variables or processes are discussed individually with the emphasis on how remote sensing is being used, and not on the technology as far as sensor details and specific instruments are concerned. More details can be found in two recent books on this general subject (Engman and Gurney, 1982; Schultz and Engman, 2000).

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

Finally, the current developments and hydrologic applications of integrated geographical information systems (GIS) technology are presented. Management and efficient utilization of large spatial data volumes is going to be one of the major challenges of the coming decades. GIS have the capability to efficiently store, manipulate, retrieve, and analyze spatially referenced data. This is the primary reason why GIS are becoming popular among the hydrological community to develop new types of hydrological models and to modify existing models to incorporate widely available spatial data.

2 PRECIPITATION

Recognizing the practical limitations of rain gages for measuring spatially averaged rainfall over large areas and inaccessible areas, hydrologists have increasingly turned to remote sensing as a means for quantifying the precipitation input, especially in areas where there are few surface gages. Because the fundamental approach to measuring rainfall and snow are different with respect to remote sensing, snow is discussed separately.

Direct measurement of rainfall from satellites for operational purposes has not been generally feasible because the presence of clouds prevents observation of the precipitation directly with visible, near-infrared and thermal-infrared sensors. However, improved analysis of rainfall can be achieved by combining satellite and conventional gage data. Satellite data are most useful in providing information on the spatial distribution of potential rain-producing clouds, and gage data are most useful for accurate point measurements. Although ground-based radar, which is a remote-sensing technique, has advanced to an operational stage for locating regions of heavy rain and for estimating rainfall rates, it will not be discussed in this chapter.

Useful data can be derived from satellites used primarily for meteorological purposes, including polar orbiters such as the NOAA-N series and the Defense Meteorological Satellite Program, and from geostationary satellites such as *GOES* (Geostationary Orbiting Environmental Satellite), *GMS*, and *Meteosat*. However, their visible and infrared images can provide information only about the cloud tops rather than cloud bases or interiors. Since these satellites provide frequent observations (even at night with thermal sensors), the characteristics of potentially precipitating clouds and the rates of changes in cloud area and shape can be observed. From these observations, estimates of rainfall can be made that relate cloud characteristics to instantaneous rainfall rates and cumulative rainfall over time. For example, Strubing and Schultz (1983) have developed a runoff regression model that is based on Barrett's (1970) indexing technique. The cloud area and temperature are the satellite variables used to develop a temperature-weighted cloud cover index. This index is then transformed linearly to mean monthly runoff. Rott et al. (1986) also developed a daily runoff model using *Meteosat* data for a cloud index. Schultz (1994) has demonstrated the use of the infrared channel from *Meteosat* to estimate monthly rainfall volumes using a modified Arkin approach (Papadakis et al., 1992). The monthly rainfall data were transformed

into monthly runoff volumes for the 16,000 km^2 Tano basin in West Africa using a model based on a series of nonlinear reservoirs. The results were reasonably good and certainly adequate for water resources planning. For the practicing hydrologist, satellite rainfall methods are most valuable when there are no or very few surface gages for measuring rainfall.

Tsonis et al. (1996) investigated the ability of visible and infrared satellite data to produce rainfall estimates for input to the National Weather Service river forecast model that had been calibrated with rain gage data. They found good correlations with gage data for areas over 10,000 km^2. In a companion study, Guetter et al. (1996), using the satellite-derived rainfall estimates produced streamflow and soil moisture estimates using the river forecast model. They concluded that flow simulation accuracy is sensitive to basin scale with better results being produced from larger basins. Derived soil moisture estimates were similar to those simulated with gaged data for the surface layer but lower for the deep soil moisture.

3 SNOW HYDROLOGY

Snow is a form of precipitation; in hydrology it is treated somewhat differently because of the lag between when it falls and when it produces runoff and groundwater recharge, and is involved in other hydrologic processes. Remote sensing is a valuable tool for obtaining snow data for predicting snowmelt runoff as well as climate studies. Nearly all regions of the electromagnetic spectrum provide useful information about the snowpack. Depending on the need, one may like to know the areal extent of the snow, its water equivalent, or the "condition" or grain size, density, and presence of liquid water within the snowpack. Although no single region of the spectrum provides all these properties, techniques have been developed to provide all of the properties to some degree or other.

The water content of snow can be measured from low elevation aircraft carrying sensitive γ-radiation sensors. This approach is limited to low aircraft altitudes (approximately 150 m) because the atmosphere attenuates a significant portion of the γ radiation. Currently, this operational program covers over 1400 flight lines annually in the United States and Canada. This method is effective for measuring snow cover in open plains, but is less effective in more hilly terrain or when there is extensive forest cover. Use of satellite data for snow mapping has become operational in several regions of the world. Currently, the National Oceanic and Atmospheric Administration (NOAA) develops snow cover maps for about 3000 river basins in North America of which approximately 300 are mapped according to elevation for use in streamflow forecasting (Carroll, 1990). NOAA also produces regional and global maps of mean monthly snow cover.

Microwave remote sensing offers great promise for future applications to snow hydrology. This is because the microwave data can provide information on the snowpack properties of most interest to hydrologists, i.e., snow cover area, snow water equivalent (or depth), and the presence of liquid water in the snowpack, which signals the onset of melt. With the availability of satellite microwave data Scanning

Multichannel Microwave Radiometer (SSMR) and Special Sensor Microwave/ Imager (SSM/I), algorithms have been developed for estimating snow water equivalent for dry snow and mapping the depth and global extent of snow cover (Chang et al., 1987). The passive microwave systems are limited by their interaction with other media such as forest areas, although a method to correct for the absorption of the snow signal by forest cover has been developed (Chang et al., 1991). The spatial resolution attainable by the passive satellite systems is also a limitation but Rango et al. (1989) have shown that that reasonable snow water equivalent estimates can be made on basins smaller than 10,000 km^2.

Active microwave remote sensing also has the potential to provide important information about the snowpack at very high resolution with synthetic aperture radar (SAR) (Stiles et al., 1981; Rott, 1986). Unfortunately, analysis of radar data is more complex than passive microwave data and, until very recently, there have been no orbiting SAR systems for collecting snow data. In spite of that, aircraft and shuttle SAR measurements have shown that SAR can discriminate between snow and glaciers from other targets and discriminate between wet and dry snow (Shi and Dozier, 1992, 1995).

Snowmelt runoff procedures that use remote sensing can be grouped into empirical approaches and modeling. Early use of remote sensing focused on empirical relationships between snow cover area or percent snow cover and monthly or accumulated runoff. These simple relationships work very well for some applications, particularly in data-sparse regions of the world. The snowmelt runoff model (SRM) (Martinec et al., 1983) was specifically developed for using remote sensing of snow cover by elevation zone as the primary input variable. Although SRM uses a simple degree-day melt model, it applies the model to the different elevation zones to account for the areal distribution of the snow. SRM has been extensively tested on basins of different sizes and regions of the world. Although SRM is a degree-day model that uses only snow cover as remote-sensing derived input, this model has been recently modified to include a simple snowmelt energy budget algorithm (Kustas et al., 1994). This model has been tested against lysimeter data and suggests that the radiation-based snowmelt factor may improve runoff predictions at the basin scale.

4 SOIL MOISTURE

Recent advances in remote sensing have shown that soil moisture can be measured by a variety of techniques. However, only microwave technology has demonstrated a quantitative ability to measure soil moisture under a variety of topographic and vegetation cover conditions so that it could be extended to routine measurements from a satellite system.

The major factor inhibiting widespread use of remotely sensed soil moisture data in hydrology is the lack of data sets and optimal satellite systems. For the most part, scientists have been restricted to data from short-duration aircraft campaigns or analysis of the SMMR and SSM/I passive microwave satellites. Although the avail-

able passive systems do not have the optimum wave lengths for soil moisture, research has demonstrated that in areas of sparse vegetation a valuable estimate can be obtained (Owe et al., 1988). Historical data from the SSMR passive microwave system is more valuable than the SSM/I data because it had a C-band radiometer, which is a better instrument for soil moisture (Owe et al., 1992); however, its period of record is limited to 1982 to 1987. In both cases the footprint is rather large, varying from about 25 km for the SSM/I to about 150 km for the C-band SMMR.

The SAR systems offer perhaps the best opportunity to measure soil moisture routinely over the next few years. Currently, the European Resources Satellite (*ERS-1*) C-band and Japan Environmental Resources Satellite (*JERS-1*) L-band SARs and the Canadian *RADARSAT* (also C-band) are operational. Although it is believed that an L-band system would be optimum for soil moisture, the preliminary results from the *ERS-1* C-band radar demonstrate its capability as a soil moisture instrument. One main drawback to the existing SAR systems is that there are no existing algorithms for the routine determination of soil moisture from single-frequency, single-polarization radars. A second limitation comes from their long period between repeat passes; for the most part it is 35 to 46 days, although the *RADARSAT* has 3-day capability for much of the globe in a SCANSAR (wide swath, 500 km) mode.

There continues to be speculation about the potential value for soil moisture as an input variable in hydrologic models, either to establish the initial conditions for simulating storm runoff or as a descriptor of hydrologic processes. To date there has been more promise than substance, but initial progress is beginning to appear as some of the aircraft experimental data become available.

Aircraft data taken during the Fist ISLSCP Field Experiment (FIFE) campaign were used to map the spatial pattern of soil moisture resulting from drainage and ET in a 37.7-ha watershed (Wang et al., 1989). These patterns, shown in Figure 1, were seen to match the results of a simple slab model and identified the region contributing base flow to the channel (Engman et al., 1989). Attempts to use passive microwave measurements in a small watershed showed good correlation with the ground data and may yield a reliable technique for calibrating the model (Wood et al., 1993). Even the relatively low-resolution passive data can improve the water budget calculations of a small basin (Lin et al., 1994). Goodrich et al. (1994) studied the prestorm soil moisture at various scales of basin runoff. They concluded that initial values were important but that the resolution of the final remote-sensing product was not a limitation.

The value of remotely sensed soil moisture data in a semidistributed hydrology model was demonstrated using data from the 1992 Washita microwave experiment. Initializing the surface soil moisture fields with the Electronically Steered Thinned Array Radiometer (ESTAR) L-band microwave data produced more accurate model predictions of soil moisture changes and absolute values than those produced from the model initialized with streamflow data (O'Neill and Hsu, 1997).

The feasibility of synthesizing distributed fields of remotely sensed soil moisture by the four-dimensional data assimilation applied to a hydrological model, TOPLATS, has been explored (Houser et al., 1998) with several alternative assimilation schemes. The synthetic soil moisture fields were assimilated from remote-

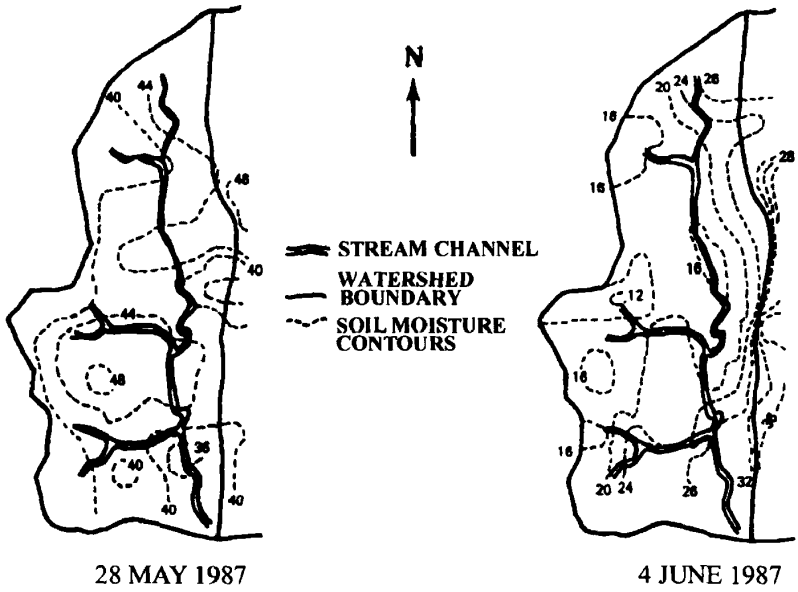


Figure 1 Temporal and spatial patterns of soil moisture in a small drainage basin illustrating the drying pattern (after Wang et al., 1989).

sensing soil moisture data and the output of a soil–vegetation–atmosphere scheme. The spatially distributed hydrology model's descriptive ability was improved with the assimilation of the soil moisture data.

5 EVAPOTRANSPIRATION

In general, remote-sensing techniques cannot measure evaporation or evapotranspiration directly. However, remote sensing does have two potentially very important roles in estimating evapotranspiration. First, remotely sensed measurements offer methods for extending point measurements or empirical relationships, such as the Thornthwaite (1948), Penman (1948), and Jensen and Haise (1963) methods, to much larger areas, including those areas where measured meteorological data may be sparse. Secondly, remotely sensed measurements may be used to measure variables in the energy and moisture balance models of evapotranspiration. Although there has been progress made in the direct remote sensing of the atmospheric parameters that affect evapotranspiration, such as the Rahman LIDAR, this is essentially a ground-based, point measurement and will not be covered in this report.

The question of how to use the spatial nature of remote-sensing data to extrapolate point evapotranspiration measurements to a more regional scale has been addressed in several ways. Using the temperature sounders on the meteorological

satellites in a linear regression model, Davis and Tarpley (1983) estimated shelter temperatures with an error of about 2 K for clear or partly cloudy conditions. Price (1982) used thermal data from the Heat Capacity Mapping Mission (HCMM) to estimate regional-scale evapotranspiration rates, which were found to be comparable to pan evaporation data. Jackson (1985) and Gash (1987) have proposed an analytical framework for relating the horizontal changes in evaporation to horizontal changes in surface temperature. Kustas et al. (1990) demonstrated these concepts for an agricultural area under clear sky conditions. Humes et al. (1994) has proposed a simple model using remotely sensed surface temperatures and reflectances for extrapolating energy fluxes from a point to a regional scale; however, other than for clear sky conditions, variations in incoming solar radiation, meteorological conditions, and surface roughness limit this approach.

Several variables related to the energy balance equation can be measured by remote sensing and simple meteorological measurements. Generally, the latent heat term is determined as the residual of the other terms in the energy balance. Incoming solar radiation can be estimated from satellite observations of cloud cover, primarily from geosynchronous satellites (Brakke and Kanemasu, 1981; Tarpley, 1979). Pinker and Laszlo (1992) have proposed a model that infers incoming short-wave fluxes and surface albedos from *GOES* data. Pinker et al. (1994) used this model to demonstrate that incoming shortwave radiation can be measured quite accurately, even under variable cloud conditions, at the basin scale.

For clear sky conditions, the surface albedo may be estimated by measurements covering the entire visible and near-infrared waveband, while empirical relations using narrow spectral bands can be used to determine albedo over heterogeneous surfaces (Jackson, 1985; Brest and Goward, 1987). Although albedo estimated this way is not the true hemispherical albedo, lack of directional data or simple models make this correction not feasible under most applications.

Surface temperature can be estimated from measurements in thermal infrared wavelengths, that is, the 10.5- to 12.5-μm waveband, either assuming a surface emissivity (close to unity for natural surfaces) or having measured values of the surface emissivity. Surface temperatures can be used to estimate the outgoing long-wave radiation term in the net radiation equation (Kustas et al., 1994).

The soil heat flux term can be estimated with remote-sensing measurements. A simplified approach defines the ratio of soil heat flux to net radiation in terms of vegetation cover, which, in turn, is determined from visible and near-infrared measurements (Clothier et al., 1986; Choudhury et al., 1987; Kustas and Daughtry, 1990). The diurnal effects (Owe and van de Grind, 1990) and influence of soil moisture (Brutsaert, 1982) are assumed to be secondary for large areas (Kustas et al., 1994).

The sensible heat flux can be estimated using several approaches, including the bulk resistance approach proposed by Monteith (1973) and similarity principles for the unstable boundary layer (Brutsaert and Sugita, 1992), where the surface temperatures are measured by remote sensing. These approaches have met with varying degrees of success (Hall et al., 1992; Brutsaert and Sugita, 1992; Brutsaert et al., 1993; Kustas et al., 1994).

Additional approaches for estimating ET from remote sensing data are being explored. Ottle et al. (1989) have shown how satellite-derived surface temperatures can be used to estimate ET and soil moisture in a model that has been modified to use these data. Mauser (1990) has shown how multitemporal Système Probatoire pour l'Observation de la Terre (*SPOT*) and Thematic Mapper (TM) data to derive plant parameters for estimating ET in a GIS-based model. Later, Mauser (1996) used Advanced Very High Resolution Radiometer (AVHRR) thermal data to validate an actual ET mesoscale model by comparing them to the surface temperature distributions. Soares et al. (1988) demonstrated how thermal infrared and C-band radar could be used to estimate bare soil evaporation. Choudhury et al. (1994) have shown strong relationships between evaporation coefficients and vegetative indices.

6 RUNOFF

One of the first applications of remote-sensing data in hydrologic models used *Landsat* data to determine both urban and rural land use for estimating runoff coefficients (Jackson et al., 1976). Land use is an important characteristic of the runoff process that affects infiltration, erosion, and evapotranspiration. Distributed models, in particular, need specific data on land use and its location within the basin. Most of the work on adapting remote sensing to hydrologic modeling has involved the Soil Conservation Service (SCS) runoff curve number model (U.S. Department of Agriculture, 1972) for which remote-sensing data are used as a substitute for land cover maps obtained by conventional means (Jackson et al., 1977; Bondelid et al., 1982).

In remote-sensing applications, one seldom duplicates detailed land-use statistics exactly. For example, a study by the Corps of Engineers (Rango et al., 1983) estimated that an individual pixel may be incorrectly classified about one-third of the time. However, by aggregating land use over a significant area, the misclassification of land use can be reduced to about 2%, which is too small to affect the runoff coefficient or the resulting flood statistics.

Studies have shown (Jackson et al., 1977) that for planning studies the *Landsat* approach is cost effective. The authors estimated that the cost benefits were on the order to 2.5 to 1 and can be as high as 6 to 1, in favor of the *Landsat* approach. These benefits increase for larger basins or for multiple basins in the same general hydrological area. Mettel et al. (1994) demonstrated the recomputation of Probable Maximum Flood (PMFs) for the Au Sable River using HEC-1 and updated and detailed land-use data from Landsat TM resulted in 90% cost cuts in upgrading dams and spillways in the basin.

7 WATER AND ENERGY BALANCE MODELS

In recent work, Dubayah and Lettenmaier (1997) have attempted to maximize the use of remote-sensing data as drivers for a large-scale coupled water and energy balance model. They used the VIC-3L model (Liang et al., 1994) applied to the

Arkansas–Red River basins in the Southern Great Plains in the United States. There were two objectives to this research: (1) to develop and test a land surface hydrologic model capable of using remote-sensing data and (2) to develop and test algorithms for generating data from remote-sensing measurements. Remote-sensing data were obtained from *GOES* (solar radiation), AVHRR (downwelling long-wave radiation, air temperature, surface humidity, and vegetation), Normalized Difference Vegetation Index (NDVI), Leaf Area Index (LAI) (canopy interception, and canopy resistance).

The VIC-3L model was used to simulate the water and energy fluxes for the month of June, 1987. The model was first used with ground-based data (from 26 meteorological stations) and then with the remote-sensing-derived data. Comparison of the results yielded differences of as much as 40% in net radiation, 15% in latent heat, and 100% in sensible heat. For such studies, the results from the ground-based data are not necessarily correct; it is not known whether the remote-sensing or the ground-based data give the correct results. Thus it really could not be determined if the remote-sensing data resulted in an improvement or a degradation in the water and energy fluxes. However, the remote-sensing simulations did provide a spatial pattern that appears to provide more information about the distribution of the fluxes than do the ground-based measurements.

8 GEOGRAPHICAL INFORMATION SYSTEMS

Geographical information systems provide appropriate methods for efficient storage, retrieval, manipulation, analysis, and display of large volumes of spatially referenced data. Accordingly, GIS consist of four basic components: data input and editing, storage of geographic databases, data analysis and spatial modeling, and data visualization and presentation (Fig. 2). The data may be collected from fieldwork, extraction of map data, air photo interpretation, and interpretation and classification of remotely sensed images. Data input may be carried out by manual digitization or computer-assisted semiautomatic methods. The data are organized into a series of spatially co-registered layers, with each layer relating to a particular theme or a set of layers relating to temporal variation of a theme.

Data input and structuring is one of the most time-consuming and expensive tasks in the creation of a GIS. Remotely sensed data can be put to the best use if they are incorporated in GIS. A GIS, therefore, when combined with up-to-date data from a remote-sensing system, can assist in the automation of several operations (e.g., interpretation, change detection, map revisions). The hydrological system is a dynamic entity; the information stored in a GIS is only a static representation of the real world and therefore has to be updated for the temporal coverage (i.e., a third dimension) on a regular basis. Remotely sensed satellite data offer an excellent input in this context to provide repetitive, synoptic, and accurate information of the changes of a watershed.

Integration of GIS with hydrological models is necessary to better explain the complexity of hydrological processes arising from spatial heterogeneity of inPut

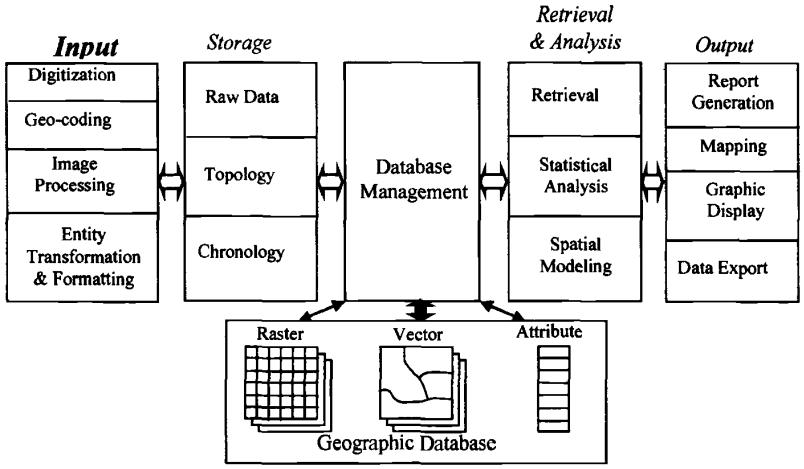


Figure 2 Submodules of a GIS for input, storage, retrieval, analysis, modeling, and presentation of spatial and nonspatial data.

parameters such as topography, soil types and characteristics, vegetation, and antecedent soil conditions. A GIS can be employed to supply physical and hydrological parameters to a hydrological model. Model simulation results can be analyzed using a GIS. This requires a continuous flow of information to and from both the GIS and hydrological model. One of the useful possibilities of linking GIS with hydrological modeling is the capability of dynamic spatial visualization of the model simulation results, including user interaction. Real-time or near-real-time visualization of simulated hydrological processes could greatly improve existing analysis of simulation results.

Integration of Remotely Sensed Data into GIS

Remote sensing can be incorporated into the system in a variety of ways: as a measure of land use, impervious surfaces, for providing initial conditions for flood forecasting, and for monitoring flooded areas (Neumann et al., 1990). The GIS allows for the combining of other spatial data forms such as topography, soils maps as hydrologic variables such as rainfall distributions, or soil moisture. This approach was demonstrated by Kouwen et al. (1993) where their grouped response unit (GRU) included satellite-based land use and lies within a computational element that may be either a sub-basin or and area of uniform meteorological forcing. In HYDROTEL, Fortin and Bernier (1991) propose combining *SPOT* DEM (digital elevation model) data with satellite-derived land use and soils mapping data to define homogeneous hydrologic units (HHU). In a study of the impact of land-use change on the Mosel River Basin, Ott et al. (1991) and Schultz (1993) have defined hydrologically similar units (HSU) by DEM data, soils maps, and satellite-derived land

use. They also used satellite data to determine a vegetation index (NDVI) and a leaf water content index (WCI), which are combined to delineate areas where a subsurface supply of water is available to vegetation. The distribution of microwave remotely sensed near-surface (0–5-cm deep) soil moisture was analyzed to identify areas of high soil moisture gradients (Mattikalli et al., 1998). This analysis showed a direct correlation between soil moisture dynamics and soil texture. Soil moisture data were employed in a hydrological model linked to a GIS, to predict subsurface hydraulic conductivity.

Remote-sensing systems use raster format for collection and acquisition of data. Many of the commonly employed GIS systems (e.g., Arc/Info) mainly use the vector format to store data layers. In this format, data are collected as points, lines, and polygons, where each structure holds information for a specific region (Fig. 3). Both the vector and raster structures have advantages and difficulties that are well described in the literature (e.g., Peuquet, 1984; Burrough, 1990), yet their fundamental differences make the integration a complicated task (Piwowar et al., 1990). In the recent past, many commercial GIS have been adapted to offer raster image display and handling capabilities (e.g., Arc/Info Version 6.0 or later), and several others offer both raster and vector capabilities (e.g., GRASS). The integration of remotely sensed data with GIS data occurs naturally in a raster GIS because data structures are approximately the same for both sources. In a vector system, the integration requires more effort, and several technical problems need to be overcome for the true integration. Important problems in the integration are the raster/vector dichotomy, generalization, and accuracy of digital information (Piwowar et al., 1990; Lunetta et al., 1991). Although the raster/vector dichotomy is a major impediment for a true integration, a significant advancement has been made to resolve the issue (e.g., McKeown, 1987; Conese et al., 1992; van der Laan, 1992; Westmoreland and Stow, 1992). These studies have employed a variety of approaches including use of quadtrees, object-oriented methods, knowledge-based systems, expert systems, artificial intelligence, etc. to achieve the task of true integration (Fritsch, 1992; Molenaar and Janssen, 1992). Examples of some commercially available systems

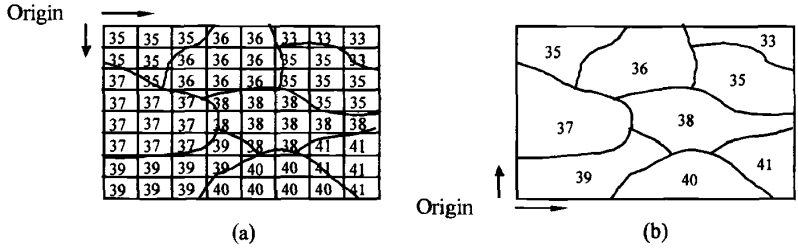


Figure 3 Representation of spatial data in a GIS: (a) raster-formatted data consists of a sequence of orderly placed pixels (or picture elements) and (b) vector-formatted data consists of polygon entities to represent features.

that have some integration capabilities include GRASS, Arc/Info Version 6.0 onwards, ERDAS IMAGINE, and PCI.

Integration of raster and vector data types requires an efficient raster-to-vector (and/or vice versa) conversion routine. Mattikalli et al. (1995) developed a methodology for the separate but equal type of integration, in which the key process is a raster-to-vector (and vice versa) conversion. The procedure makes use of some built-in routines commonly available in most vector GIS, and some intermediate data formats, viz. lattice and SVF (single variable file). This approach has been employed by Mattikalli (1995) to integrate remotely sensed satellite data derived from both fine- and coarse-resolution sensors with digitized map data.

GIS and Digital Terrain Models

Digital terrain modeling is one of the strong areas where GIS has been widely utilized in hydrology. Digital elevation model (DEM) data are employed to derive watershed characteristics such as slope, aspect, curvature, drainage network structure (e.g., Fairfield and Leymarie, 1991), hydrologic response units (HRUs), and also to delineate watershed boundaries (Band, 1986; Jensson and Domingue, 1988; Schultz, 1994). Lozar (1992) delineated drainage paths and watersheds of the entire Earth based on the 5-arc-minute DEM of Earth's land surface. Remotely sensed information can be integrated with DEM for a variety of hydrologic applications. For example, Dubayah (1992) employed a DEM, *Landsat* TM data, and a radiative transfer algorithm to model spatial variability of net solar radiation at fine spatial resolution. Also, remotely sensed products are employed in conjunction with a DEM to produce realistic perspective views of a watershed that aid visualization and understanding of spatial and temporal variability of hydrological parameters (Gugan and Dowman, 1988).

GIS and Hydrologic Models

Watershed database development usually is the first important stage in a hydrologic modeling study. Remotely sensed data might be employed to generate thematic maps and also to serve as map basis when no other reliable data are available. *Landsat* TM and *SPOT* images data are suitable for production of digital map at scales ranging from 1 : 50,000 to 1 : 100,000 (Welch et al., 1985; Swann et al., 1988; Gugan and Dowman, 1988; Konecny et al., 1988). Base maps, produced from remote sensing and integrated within a GIS, hold promise in terms of greater reliability, i.e., lower meta-uncertainty (uncertainty about uncertainty) for map information because errors are known and tracked throughout the map generation process. Overlaying, merging, and performing map calculations are key GIS features often used in many hydrological applications. Schultz (1993) presents an example in which soil water storage information was derived by merging plant root depth data (derived from land-use classification of *Landsat* image) and soil porosity data (derived from digitized soil maps).

Historically, runoff modeling at the river basin scale has lumped rainfall, infiltration, and other hydraulic parameters to apply everywhere in the basin. With the advent of distributed modeling, a basin is subdivided into computational elements at a smaller scale. A distributed simulation model allows a user to simulate spatially variable parameters without lumping. However, setting up such a model with spatially distributed data and parameters is a time-consuming and laborious task. If a GIS is integrated with the model, these chores become much easier and often transparent to the user. An additional advantage of integrating distributed numerical models with a GIS includes calculation and display of runoff flow depths across watershed sub-basins.

The runoff curve number (CN) approach (USDA, 1972) to rainfall–runoff modeling is appealing for an integrated remote-sensing and GIS environment. This approach estimates volume of direct runoff (Q) in terms of volume of rainfall (P) and potential maximum storage (S), which is derived from the CN, a coefficient that is directly related to watershed land use, land management, and soil properties. Since land use can be routinely monitored using remote sensing, it is possible to analyze the effects of land-use changes (e.g., urbanization) on watershed runoff. Figure 4 shows various stages of computation of this approach implemented within a GIS. Mattikalli et al. (1996) employed Arc/Info to store various input parameters as thematic layers and generated flood hydrographs in a predominantly rural watershed. This approach has also been used to generate single-event flood hydrographs and synthetic flood frequency curves (Muzik and Chang, 1993).

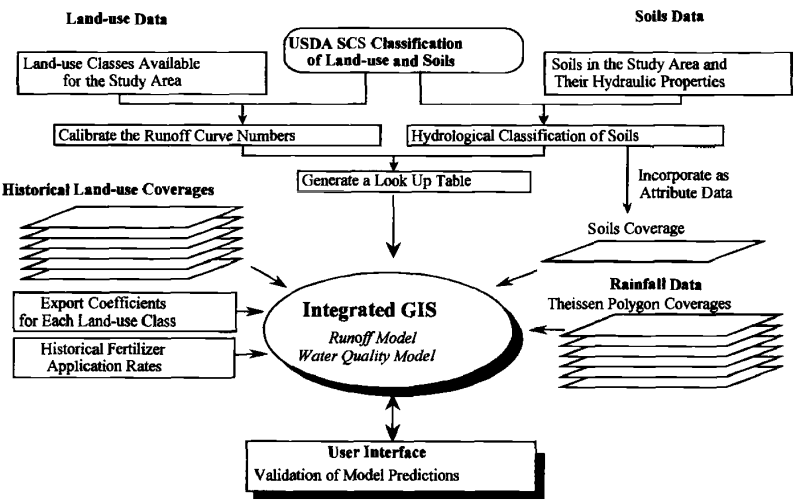


Figure 4 Schematic diagram of a GIS approach for prediction of river discharge using the SCS curve numbers and water quality using the export coefficient model (Mattikalli et al., 1996).

In urban watersheds, the spatial analysis capabilities of a GIS can be used for hydrological analysis. Watershed attributes such as soils information (infiltration rates, hydraulic conductivity, and storage capacities), surface characteristics (pervious, impervious, slope, roughness), geometry and dimensions of flow planes, routing lengths (overland, gutter, and sewer), and geometry and characteristics of routing segments can be efficiently stored and utilized for urban runoff calculations. Most of the earlier studies have used GISs to derive parameters of lumped models. For example, Johnson (1989) used GIS for the generation of input data for a digital map-based modeling system that supports lumped parameter models such as unit hydrograph, time–area, and cascade of reservoirs. The advent of distributed modeling and powerful GIS allows modelers to simulate spatially variable parameters. To date, several hydrological models such TOPMODEL and CREAMS have been integrated to operate within GIS environments (Chairat and Delleur, 1993; Romanowicz et al., 1993). Moeller (1991) used GIS to determine input parameters for the HEC-1 model, Sircar et al. (1991) used a GIS to determine time–area curves. Djokic and Maidment (1991) used Arc/Info with the rational method to determine inlet and pipe capacity of an urban storm sewer system. Kim and Ventura (1993) used a GIS to manage and manipulate the land-use data for modeling the non-point-source pollution of an urban basin using an empirical urban water quality model. Greene and Cruise (1996) employed Arc/Info GIS to derive urban watershed feature attributes (location coordinates, parameters of runoff generating polygons, gutters and storm drains) for input into hydrologic modeling procedures to estimate runoff. Vieux (1991) developed a method for modeling direct surface runoff using a combination of the finite-element method and GIS. Schultz (1994) presents three different examples on hydrological modeling using remote sensing in the framework of ILWIS and Arc/Info GIS. These examples demonstrate merging of *Landsat* TM and *Meteosat* geostationary image products and ancillary data (viz. DEM and its derived products) stored in a GIS for rainfall/runoff modeling and water balance parameter computation at 30 m, 5 km, and at HRU spatial scales. Mattikalli et al. (1996) employed the runoff curve number (CN) approach to compute direct runoff depth and its spatial and temporal variations based on historic remotely sensed data within a GIS framework.

Water quality modeling applications using remote sensing and GIS have concentrated mainly on non-point source (NPS) pollution. To date, several water quality models (AGNPS, ANSWERS, USLE, export coefficient model, etc.) have been interfaced with GIS. The spatially distributed agricultural non-point-source (AGNPS) model integrated with GIS (Srinivasan and Engel, 1994) allows modelers to handle each point source, pesticide, and channel information in a decision support system, WATERSHEDSS (Water, Soil, and Hydro-Environmental Decision Support System) (Osmond et al., 1997). Using such a system, one can determine critical areas within a watershed and evaluate effects of alternative land treatment scenarios on water quality. Mattikalli et al. (1996) implemented an export coefficient model within a vector-based GIS to quantify spatial and temporal changes of total nitrogen loading in surface water as a response to changes in watershed land use, management, and fertilizer application rates. Although this method is based on empirical

export coefficients derived from the literature, more accurate coefficients can be derived by inverse solution to a physical based model.

Management and modeling of groundwater and its quality have also been explored (e.g., Maidment, 1993; Merchant, 1994). In the majority of studies, spatial models designed to evaluate groundwater vulnerability for contamination have been implemented in GIS. However, these approaches have not employed data derived from remote sensing, probably because of the specific nature of the input parameters. The models need to be adapted to incorporate remotely sensed products and then implemented within a GIS.

Monitoring and/or prediction soil erosion computed using the universal soil loss equation (USLE) is another application of integrated GIS (e.g., Pelletier, 1985). Slope steepness (S) and slope length (L) factors are derived using DEM, and rainfall factors are assigned using the triangular irregular network (TIN) structure for the rainfall gaging stations. Erosion control practice and land-use/land-cover (or cropping management) factors are estimated using *Landsat* (Multi-Spectal Scanner (MSS) and TM) and SPOT sensor data via land-use/land-cover classification and associated land management information (Jurgens and Fander, 1993). In the revised USLE (Renard et al., 1991), the L factor has been modified for influence of profile convexity/concavity using segmentation of irregular slopes of a complex terrain. Mitasova et al. (1996) integrated regularized spline with tension for computation of S and L factors and used a unit stream power and directional derivative approach for modeling spatial distribution of areas with topographic potential for erosion or deposition.

GIS and Spatial Analysis

The synoptic nature of remote-sensing data offers an excellent opportunity to identify spatial characteristics of land surface changes and other hydrologic variables. Analysis of spatial variability is performed using different techniques including Monte Carlo methods (Fisher, 1991) and geostatistical techniques such as semivariograms and kriging (Oliver and Webster, 1990).

Three- and Four-Dimensional GIS

To date, applications have recognized the importance of change of hydrologic processes or input model parameters over time. Modern GIS have capabilities of traditional two-dimensional (2D) GIS to perform spatial analysis as well as the ability to handle and visualize third dimension (such as depth) and time as a fourth dimension (Fisher, 1993). Three-dimensional (3D) GIS are suitable for many applications in hydrology such as predictive hydrogeological modeling (Raper, 1989). Three-dimensional GIS lend themselves to the iterative process of modeling as well as the evolutionary nature of site characterization and remediation. Although most current 3D GIS provide some solutions for complex subsurface processes, they are still at the visualization stage rather than true modeling or interpretation. No one system yet meets all the needs of an ideal modeling environment,

hence integration between multiple systems is desired. Also, four-dimensional (4D) GIS do not adequately represent the temporal dimension (Langran, 1992) because no GIS currently adequately handles chronology. We typically illustrate the effects of temporal change as slices of time for discrete intervals, but we need to show dynamic change over continuous time. Although many GIS can generate a 3D diagram, no commercial system has 3D geometry and topology such that disparate databases can be integrated in three dimensions as well as they are in two dimensions. The ultimate solution would be able to handle change in time as well as change in space. An ideal GIS handling time as a fourth dimension (4D GIS) will have chronology treated much like topology; before and after taking on the same importance as left and right in 2D space or above and below in 3D space. Such 4D GIS would be of immense value for a number of research areas in hydrology including soil moisture modeling, groundwater modeling, etc. because of their inherent four-dimensional nature.

9 'ARY AND CONCLUSIONS

Continuing high spatial resolution data from the *Landsat* and *SPOT* satellites, passive microwave data from the special sensor/microwave imager (SSM/I) and continuing meteorological satellite coverage from the *NOAA*, *GOES*, *GMS*, and *Meteosat* series all mean that the remotely sensed techniques can continue to be employed and expanded upon. New sensors, particularly in the microwave region, promise great potential for hydrologic applications. There are several satellites, such as *ERS-1/2* launched by the European Space Agency, the *J-ERS-1* launched by the Japanese, and *RADARSAT* launched by the Canadians that will provide useful data for hydrologists. All carry single-polarization, single-frequency SARs. An additional satellite being planned that will have considerable hydrologic interest is the Tropical Rainfall Measurement Mission (TRMM) (Simpson et al., 1988).

The EOS (Earth Observational System) (Butler, 1988), and its counterpart European and Japanese platforms, will lead to considerable advances in the understanding of all the earth sciences, including hydrology. The EOS instruments of most interest to hydrologists would include the MODIS and AMSR; the latter is a microwave instrument with a C-band radiometer that should provide interesting data of the land surface moisture conditions. EOS also includes the organization of the data and of other earth science in an information system where time series of all the data will be readily available is also important. This data system will allow many types of data to be used simultaneously to calibrate or be assimilated into numerical models.

Future progress in the hydrological sciences will depend a great deal upon the availability of adequate data for model development and validation. Remote sensing can and should play a pivotal role in this progress. Without it, it is very possible that future progress in the hydrological sciences will be severely retarded if not completely stopped. With it, hydrological sciences should be able to advance rapidly and to successfully address some of the previously intractable problems. An issue that must be addressed is the modification or development of new models to specifically use

remote-sensing data. For the most part, existing models have not been developed to effectively use remote-sensing data. A second point is that ground-based data are frequently available at shorter time intervals than remote-sensing data. This becomes important when simulating processes that are driven by the diurnal cycle.

Another very important issue that needs to be addressed by the hydrologic and remote-sensing communities is validation. We should not automatically assume that ground-based measurements provide the "truth." Ground-based data have an inherent weakness in that they are point measurements being applied to large, inhomogeneous areas. There is a need to develop innovative approaches to validate not only the remote-sensing-derived products but also the application of water and energy balance models to large regions.

While remote-sensing systems generate large volumes of valuable spatial data, GIS offer an appropriate technology not only for efficient storage and retrieval of spatially referenced data but also for data manipulation and spatial analysis required in distributed hydrologic modeling. With the advent of an EOS suite of platforms and sensors, it is expected that the volume of data being received will require the use of fully integrated spatial information systems supported by knowledge-based techniques in all facets of data handling.

Future development of GIS application is controlled by the state of technology, and therefore assessing the probable developments is a difficult task. At the present time, the current level of integrated applications utilizes an environment largely free of the logistical considerations of data transfer between remote sensing and GIS. Over the next few years, more efforts need to be focused on the fundamental aspects of integration such as data generalization and accuracy specification. Several problems of a true integration of remote sensing and GIS could probably be solved by recognition that GIS and remote-sensing systems process and manage spatial information at different levels of representation. Ultimately, GIS and remote sensing should be viewed as one entity that will be concerned with handling and analyzing spatial hydrologic data. The unification of these technologies will lead to a synergistic integration of spatial data handling, and the final system would have more application capabilities than just the sum of the two.

REFERENCES

Band, L. E., Topographic partition of watersheds with digital elevation models, *Water Resour. Res.*, *22*(1), 15–24, 1986.

Barrett, E. C., The estimation of monthly rainfall from satellite data, *Monthly Weather Rev.*, *98*, 322–327, 1970.

Bondelid, T. R., T. J. Jackson, and R. H. McCuen, Estimating runoff curve numbers using remote sensing data, in *Proceedings of the International Symposium on Rainfall-Runoff Modeling, Applied Modeling in Catchment Hydrology*, Water Resources Publications, Littleton, CO, 1982, pp. 519–528.

Brakke, T. W., and E. T. Kanemasu, Insolation estimation from satellite measurements of reflected radiation, *Remote Sensing Environ.*, *11*, 157–167, 1981.

Brest, C. L., and S. N. Goward, Deriving surface albedo measurements from narrow band satellite data, *Int. J. Remote Sensing*, *8*, 351–367, 1987.

Brutsaert, W. H., *Evaporation into the Atmosphere: Theory, History and Application*, Reidel, Boston, MA, 1982.

Brutsaert, W. H., and M. Sugita, Regional surface fluxes from satellite-derived surface temperatures (AVHRR) and radiosonde profiles, *Boundary-layer Meteorol.*, *58*, 355–366, 1992.

Brutsaert, W. H., A. Y. Hsu, and T. J. Schmugge, Parameterization of surface heat fluxes above forest with satellite thermal sensing and boundary-layer soundings, *J. Appl. Meteorol.*, *32*(5), 910–917, 1993.

Burrough, P. A., *Principles of Geographical Information Systems for Land Resources Assessment*, Clarendon, Oxford, 1990.

Butler, D., *From Pattern to Process: The Strategy of the Earth Observing System*, NASA, Washington, DC, 1988.

Carroll, T. R., Airborne and satellite data used to map snow cover operationally in the U.S. and Canada, in *Proceedings of the International Symposium on Remote Sensing and Water Resources*, Enschede, The Netherlands, 1990, pp. 147–155.

Conese, C., G. Maracchi, F. Maselli, M. Romani, and L. Bottai, Integration of remotely sensed data into a GIS for the assessment of land suitability, *EARSeL Adv. Remote Sensing*, *1*, 173–179, 1992.

Chang, A., J. Foster, and D. K. Hall, NIMBUS-7 derived global snow cover parameters, *Ann. Glaciol.*, *9*, 39–44, 1987.

Chang, A. T. C., J. L. Foster, and A. Rango, Utilization of surface cover composition to improve the microwave determination of snow water equivalent in a mountainous basin, *Int. J. Remote Sensing*, *12*, 2311–2319, 1991.

Chairat, S., and J. W. Delleur, Integrating a physically based hydrological model with GRASS, in *HydroGIS 93*, IASH Publ. No. 211, 1993, pp. 143–150.

Choudhury, B. J., S. B. Idso, and R. J. Reginato, Analysis of an empirical model for soil heat flux under a growing wheat crop for estimating evaporation by an infrared-temperature based energy balance equation, *Agric. Forest Meterol.*, *39*, 283–297, 1987.

Choudhury, B. J., N. U. Ahmed, S. B. Idso, R. J. Reginato, and C. S. T. Daughtry, Relations between evaporation coefficients and vegetation indices studied by model simulations, *Rem. Sens. Environ.*, *50*(1), 1–17, 1994.

Clothier, B. E., K. L. Clawson, P. J. Pinter, Jr., M. S. Moran, R. J. Reginato, and R. D. Jackson, *Agric. Forest Meteorol.*, *37*, 75–88, 1986.

Davis, P. A., and J. D. Tarpley, Estimation of shelter temperatures from operational satellite sounder data, *J. Climatol. Appl. Meteorol.*, *22*, 369–376, 1983.

Djokic, D., and D. R. Maidment, Terrain analysis for stormwater modeling, *HP*, *5*(1), 115–124, 1991.

Dubayah, R., Estimating net solar radiation using Landsat Thematic Mapper and digital elevation data, *Water Resour. Res.*, *28*(9), 2469–2484, 1992.

Dubayah, R., and D. Lettenmaier, Combining remote sensing and hydrologic modeling for applied water and energy balance studies, paper presented at NASA EOS Interdisciplinary Working Group Meeting, San Diego, CA, 1997.

Engman, E. T., and R. J. Gurney, *Remote Sensing in Hydrology*, Chapman & Hall, London, 1982.

Engman, E. T., G. Angus, and W. P. Kustas, Relationship between the Hydrologic balance of a small watershed and remotely sensed soil moisture, in *Proceedings of the IAHS Third International Assembly*, IAHS Publ. No. 186, Baltimore, 1989, pp. 75–84.

Fairfield, J., and P. Leymarie, Drainage networks from grid digital elevation models, *Water Resourc. Res.*, *27*, 709–711, 1991.

Fisher, T. R., Integrating three-dimensional geoscientific information system (GSIS) technologies for groundwater and contaminant modeling, in *HydroGIS 93*, IAHS Publ. No. 211, 1993, pp. 235–241.

Fisher, T. R., and R. Q. Wales, Rational splines and multi-dimensional geologic modeling, in R. Pflug and J. W. Harbaugh (eds.), *Three Dimensional Computer Graphics in Modeling Geologic Struuctures and Simulating Process*, Springer-Verlag, Heidelberg, 1991, pp. 17–28.

Fortin, J.-P., and M. Bernier, Processing remotely sensed data to derive useful input data for hydrotel hydrologic model, in *Proc. IGARS*, Houston, TX, 1991.

Fritsch, D., Analysis of remote sensing data in geographical information systems, *EARSeL Adv. Remote Sensing*, *1*, 60–65, 1992.

Gash, J. H. C., An analytical framework for extrapolating evaporation measurements by remote sensing surface temperature, *Int. J. Remote Sensing*, *8*(8), 1245–1249, 1987.

Goodrich, D. C., T. J. Schmugge, T. J. Jackson, C. L. Unkrich, T. O. Keefer, R. Parry, L. B. Bach, and S. A. Amer, Runoff simulation sensitivity to remotely sensed initial soil water content, *Water Resour. Res.*, *30*(5), 1393–1405, 1994.

Greene, R. G., and J. F. Cruise, Development of a geographic information system for urban watershed analysis, *Photogrametric Engineering and Remote Sensing*, *62*(7), 863–870, 1996.

Guetter, A. K., K. P. Georgakakos, and A. A. Tsonis, Hydrologic applications of satellite data, Part 2. Flow simulations and soil water estimates, *J. Geophys. Res. Atmos.*, *101*(D21), 26527–26538, 1996.

Gugan, D. J., and I. J. Dowman, Accuracy and completeness of topographic mapping from SPOT imagery, *Photogram. Rec.*, *12*(72), 787–796, 1988.

Hall, F. G., K. F. Humerick, S. J. Goetz, P. J. Sellers, and J. E. Nickerson, Satellite remote sensing of surface energy balance: success, failures and unresolved issues, in FIFE (First ISLSCP Field Experiment) Special Issue *J. Geophys. Res.*, *97*(D17), 19061–19090, 1992.

Houser, P. R., W. J. Shuttleworth, H. V. Gupta, J. S. Famiglietti, K. H. Syed, and D. C. Goodrich, Integration of soil moisture remote sensing and hydrologic modeling using data assimilation, *Water Resour. Res.*, *34*(12), 3405–3420, 1998.

Humes, K. S., W. P. Kustas, and M. S. Moran, Use of remote sensing and reference site measurements to estimate instantaneous surface energy balance components over a semiarid rangeland watershed, *Water Resour. Res.*, *30*(5), 1363–1373, 1994.

Jackson, R. D., Estimating evapotranspiration at local and regional scales, *IEEE Trans. Geosci. Remote Sensing*, *GE-73*, 1086–1095, 1985.

Jackson, T. J., R. M. Ragan, and R. P. Shubinski, Flood frequency studies on ungaged urban watersheds using remotely sensed data, in *Proceedings of the National Symposium on Urban Hydrology, Hydraulics and Sediment Control*, University of Kentucky, Lexington, KY, 1976, pp. 31–39.

Jackson, T. J., R. M. Ragan, and W. N. Fitch, Test of Landsat-based urban hydrologic modeling, *ASCE J. Water Resourc. Planning Mgmt. Div.*, *103* (No. WR1), 141–158, 1977 (Proc. Papers 12950).

Jensen, S. K., and J. O. Domingue, Extracting topographic structure from digital elevation data for geographical information system analysis, *Photogram. Eng. Remote Sensing*, *54*, 1593–1600, 1988.

Jensen, M. E., and H. R. Haise, Estimating evapotranspiration from solar radiation, *Proc. Am. Soc. Civil Eng., J. Irrig. Drain. Div.*, *89*, 15–41, 1963.

Johnson, L. E., MAPHYD—A digital map based hydrologic modeling system, *Photogrametric Engineering and Remote Sensing*, *55*(6), 911–917, 1989.

Jurgens, C., and M. Fander, Soil erosion assessment and simulation by means of SGEOS and ancillary digital data, *Int. J. of Remote Sensing*, *14*(15), 2847–2855, 1993.

Kim, K., and S. Ventura, Large-scale modeling of urban nonpoint source pollution using a geographical information system, *Photogrametric Engineering and Remote Sensing*, *59*(10), 1539–1544, 1993.

Konecny, G., K. Jacobsen, P. Lohmann, and W. Muller, Comparison of high resolution satellite imagery, in *Proceedings of the 16th Congress of the Int. Soc. Photogrametry and Remote Sensing*, Kyoto, Japan, B9/IV, 1988, pp. 226–237.

Kouwen, N., E. D. Soulis, A. Pietroniro, J. Donald, and R. A. Harrington, Grouped response units for distributed hydrologic modeling, *J. Water Resourc. Planning Mgmt.*, *119*(3), 289–305, 1993.

Kustas, W. P., M. S. Moran, R. D. Jackson, L. W. Gay, L. F. W. Duell, K. E. Kunkel, and A. D. Matthias, Instantaneous and daily values of the surface energy balance over agricultural fields using remote sensing and a reference field in an arid environment, *Remote Sensing Environ.*, *32*, 125–141, 1990.

Kustas, W. P., M. S. Moran, K. S. Humes, D. I. Stannard, P. J. Pinter, L. E. Hipps, E. Swiatek, and D. C. Goodrich, Surface energy balance estimates at local and regional scales using optical remote sensing from an aircraft platform and atmospheric data collected over semiarid rangelends, *Water Resourc. Res.*, *30*(5), 1241–1259, 1994.

Langran, G., *Time in GIS*, Taylor and Francis, New York, 1992.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, A simple hydrologically based model on land surface water and energy fluxes for general circulation models, *J. Geophys. Res.*, *99*, 14415–14428, 1994.

Lin, D.-S., E. F. Wood, J. S. Famiglietti, and M. Mancini, Impact of microwave derived soil moisture on hydrologic simulations using a spatially distributed water balance model, in *Proceedings of the Sixth International Symposium on Physical Measurements and Signatures in Remote Sensing*, Val d'Isere, France, 1994.

Lozar, R. C., Global Climate Management by Watershed Basin Units, Construction Engineering Research Laboratory, U.S. Army Corps of Engineers, Champaign, IL, 1992.

Maidment, D. R., GIS and hydrologic modeling, in M. Goodchild, B. Parks, and L. Steyaert (Eds.), *Environmental Modeling with GIS*, Oxford University Press, New York, 1993, pp. 147–167.

Martinec, J., A. Rango, and E. Major, *The Snowmelt-Runoff Model (SRM) User's Manual*, NASA Ref. Publ. 1100, National Aeronautics and Space Administration, Washington, DC, 1983.

Mattikalli, N. M., Integration of remotely sensed raster data with vector based geographical information system for land-use change detection, *Int. J. Remote Sensing*, *16*(15), 2813–2828, 1995.

Mattikalli, N. M., B. J. Devereux, and K. S. Richards, Integration of remotely sensed satellite images with a geographical information system, *Comput. Geosci.*, *21*(8), 947–956, 1995.

Mattikalli, N. M., B. J. Devereux, and K. S. Richards, Prediction of river discharge and surface water quality using an integrated geographical information system approach, *Int. J. Remote Sensing*, *17*(4), 683–701, 1996.

Mattikalli, N. M., E. T. Engman, T. J. Jackson, and L. R. Ahuja, Microwave remote sensing of temporal variations of brightness temperature and near-surface soil water content during a watershed-scale field experiment, and its application to the estimation of soil physical properties, *Water Resour. Res.*, *34*(9), 2289–2299, 1998.

Mauser, W., Modeling the spatial variability of soil moisture and evapotranspiration with remote sensing data, in *Proceedings of the IAH Symposium on Remote Sensing and Water Resources*, Enschede, 1990, pp. 249–260.

Mauser, W., Mesoscale modeling of evapotranspiration using remote sensing data, *Proc. Europto Series*, Int. Soc. for Photo. Optical Engineering Vol. 2959, Taormina, Italy, 1996, pp. 108–117.

McKeown, D., The role of artificial intelligence in the integration of remotely sensed data with geographic information systems, *IEEE Trans. Geosci. Remote Sensing*, *25*, 330–348, 1987.

Merchant, J. W., GIS-based groundwater pollution hazard assessment: a critical review of the DRASTIC model, *Photogrametric Engineering and Remote Sensing*, *60*(9), 1117–1127, 1994.

Mettel, C., D. McGraw, and S. Strater, Money saving model, *Civil Eng.*, *64*(1), 55–56, 1994.

Mitasova, H., J. Hofierka, M. Zlocha, and L. Iverson, Modeling topographic potential for erosion and deposition using GIS, *Int. J. of Geographical Information Systems*, *10*(5), 629–641, 1996.

Moeller, R. A., Application of a geographic information system to hydrologic modeling using HEC-1, in D. B. Stafford (Ed.), *Civil Engineering Applications of Remote Sensing and GIS*, American Society of Civil Engineers, 1991, pp. 269–277.

Molenaar, M., and L. L. F. Janssen, Integrated processing of remotely sensed and geographic data for land inventory purposes, *EARSeL Adv. Remote Sensing*, *1*, 113–121, 1992.

Monteith, J. L., *Principles of Environmental Physics*, Edward Arnold, London, 1973.

Muzik, I., and C. Chang, *Flood Simulation Assisted by a GIS*, Int. Assoc. Hydrological Sciences Publication No. 211, 1993, pp. 531–539.

Neumann, P., W. Fett, and G. A. Schultz, A Geographic Information System as data base for distributed hydrological models, in *Proceedings of the International Symposium on Remote Sensing and Water Resources*, Enschede, The Netherlands, August 1990, pp. 781–791.

O'Neill, P. E., and A. Y. Hsu, The impact of microwave-derived surface soil moisture on watershed hydrological modeling, in *1997 Research and Technology Report*, NASA/GSFC, 1997.

Osmond, D. L., R. W. Gannon, J. A. Gale, D. E. Line, C. B. Knott, K. A. Phillips, M. H. Turner, M. A. Foster, D. E. Lehning, S. W. Coffey, and J. Spooner, WATERSHEDSS: A decision support system for watershed-scale nonpoint source water quality problems, *J. Am. Water Resour. Assoc.*, *33*(2), 327–341, 1997.

Ott, M., Z. Su, A. H. Schumann, and G. A. Schultz, Development of a distributed hydrological model for flood forecasting and impact assessment of landuse change in the international Mosel basin, Int. Assoc. Hydrological Sciences Publication No. 201, 1991.

Ottle, C. D., D. Vidal-Madjar, and G. Girard, Remote sensing applications to hydrological modeling, *J. Hydrol.*, *105*, 369–384, 1989.

Owe, M., and A. A. van de Grind, Daily surface moisture model for large area semi-arid land application with limited climate data, *J. Hydrol.*, *121*, 119–132, 1990.

Owe, M., A. T. C. Chang, and R. E. Golus, Estimating surface soil moisture from satellite microwave measurements and satellite derived vegetation index, *Remote Sensing Environ.*, *24*, 331–345, 1988.

Owe, M., A. A. van de Grind, and A. T. C. Chang, Surface soil moisture and satellite microwave observations in semiarid southern Africa, *Water Resour. Res.*, *28*(3), 829–839, 1992.

Papadakis, I., J. Napiorkowski, and G. A. Schultz, Monthly runoff generation by nonlinear models using multi-spectral and multi-temporal images, in *Remote Sensing Oceanogr. Hydrol. Agric. Adv. Space Res.*, *13*(5), 181–186, 1992.

Pelletier, R. E., Evaluating non-point pollution using remotely sensed data in soil erosion models, *Journal of Soil and Water Conservation*, *40*, 332–335, 1985.

Penman, H. L., Natural evaporation from open water, bare soil and grass, *Proc. R. Soc. A*, *193*, 129–145, 1948.

Peuquet, D. J., A conceptual framework and comparison of spatial data models, *Cartographica*, *21*(4), 66–113, 1984.

Pinker, R. T., and I. Laszlo, Modeling surface solar irradiation for satellite applications on global scale, *J. Appl. Meteorol.*, *31*, 194–211, 1992.

Pinker, R. T., W. P. Kustas, I. Laszlo, M. S. Moran, and A. R. Huete, Basin-scale irradiance estimates in semi-arid regions using GOES-7, *Water Resourc. Res.*, *30*(5), 1375–1386, 1994.

Piwowar, J. M., E. F. LeDrew, and D. J. Dudycha, Integration of spatial data in vector and raster formats in a geographic information system environment, *Int. J. Geogr. Inf. Syst.*, *4*, 429–444, 1990.

Price, J. C., Estimation of regional scale evapotranspiration through analysis of satellite thermal-infrared data, *IEEE Trans. Geosci. Remote Sensing*, *GE-20*, 286–292, 1982.

Rango, A., A. Feldman, T. S. George III, and R. M. Ragan, Effective use of Landsat data in hydrologic models, *Water Resour. Bull.*, *19*, 165–174, 1983.

Rango, A., J. Martinec, A. T. C. Chang, J. L. Foster, and V. F. van Katwijk, Average areal water content of snow in a mountainous basin using microwave and visible satellite data, *IEEE Trans. Geosci. Remote Sensing*, *27*, 740–745, 1989.

Raper, J. F., The three-dimensional geoscientific mapping and modeling system: A concept design, in J. F. Raper (Ed.), *Three Dimensional Applications in Geographic Information Systems*, Taylor and Francis, London, 1989, 11–19.

Renard, G. K., G. R. Foster, G. A. Weesies, and J. P. Porter, RUSLE—Revised universal soil loss equation, *Journal of Soil and Water Conservation*, *46*, 30–33, 1991.

Romanowicz, R., K. Beven, and J. Freer, *TOPMODEL as an Application Module within WIS*, IAHS Publication No. 211, 1993, pp. 211–223.

Rott, H., *Prospects of Active Remote Sensing for Snow Hydrology, Hydrologic Applications of Space Technology*, IAHS Publ. No. 160, 1986, pp. 215–233.

Rott, H., J. Aschbacher, and K. G. Lenhart, Study of river runoff prediction based on satellite data, European Space Agency Final Report No. 5376, 1986.

Shi, J. C., and J. Dozier, Radar response to snow wetness, in *Proc. International Geoscience and Remote Sensing 92*, CH3041-1, 1992, pp. 927–929.

Shi, J. C., and J. Dozier, Inferring snow wetness using C-band data from SIR-C's polarimetric synthetic aperture radar, *IEEE Trans. Geosci. Remote Sensing*, *33*(4), 905–914, 1995.

Schultz, G. A., Hydrological modeling based on remote sensing information, *Adv. Space Res.*, *13*(5), 149–166, 1993.

Schultz, G. A., Meso-scale modeling of runoff and water balances using remote sensing and other GIS data, *Hydrological Sciences Journal*, *39*(2), 121–142, 1994.

Schultz, G. A., and E. T. Engman (Eds.), *Remote Sensing in Hydrology and Water Management*, Springer, Berlin, 2000.

Simpson, J., R. F. Adler, and G. R. North, A proposed Tropical Rainfall Measuring Mission (TRMM) satellite, *Bull. Am. Meteorol. Soc.*, *69*, 278–295, 1988.

Sircar, J. K., R. M. Ragan, E. T. Engman, and R. A. Fink, A GIS based geomorphic approach for the computation of time-area curves. in D. B. Stafford (Ed.), *Civil Engineering Applications of Remote Sensing and GIS*, American Society of Chemical Engineers, New York, 1991, pp. 287–296.

Soares, J. V., R. Bernard, O. Toconet, D. Vidal-Madjar, and A. Weill, Estimation of bare soil evaporation from microwave measurements, *J. Hydrol.*, *99*, 281–296, 1988.

Srinivasan, R., and B. A. Engel, A spatial decision support system for assessing agricultural nonpoint source pollution, *WRB*, *30*(3), 441–462, 1994.

Stiles, W. H., F. T. Ulaby, and A. Rango, Microwave measurements of snowpack properties, *Nordic Hydrol.*, *12*, 143–166, 1981.

Strubing, G., and A. Schultz, Estimation of monthly river runoff data on the basis of satellite imagery, in *Proc. Hamburg Symposium*, Int. Assoc. Hydrological Sciences Publication No. 145, 1983, pp. 491–498.

Swann, R., D. Hawkins, A. Westwell-Roper, and W. Johnstone, The potential for automated mapping from geocoded digital image data, *Photogram. Eng. Remote Sensing*, *54*(2), 187–193, 1988.

Tarpley, J. D., Estimating incident solar radiation at the surface from geostationary satellite data, *J. Appl. Meteorol.*, *18*, 1172–1181, 1979.

Thornthwaite, C. W., An approach toward a rational classification of climates, *Geophys. Rev.*, *38*, 55–94, 1948.

Tsonis, A. A., G. N. Triantafyllou, and K. P. Georgakakos, Hydrologic applications of satellite data for rainfall estimation, *J. Geophys. Res. Atmos.*, *101*(D21), 26517–26525, 1996.

U.S. Department of Agriculture. Soil conservation service, in *National Engineering Handbook, Section 4, HYDROLOGY*, U.S. Government Printing Office, Washington, DC, 1972.

van der Laan, F. B., Integration of remote sensing in a raster and vector GIS environment, *EARSeL Adv. Remote Sensing*, *1*, 71–80, 1992.

Vieux, B. E., Geographic information systems and non-point source water quality and quantity modelling, *Hydrol. Process.*, *5*, 101d 113, 1991.

Wang, J. R., J. C. Shiue, T. J. Schmugge, and E. T. Engman, Mapping soil moisture with L-band radiometric measurements, *Remote Sensing Environ.*, *27*, 305–312, 1989.

Welch, R., T. R. Jordan, and M. Ehlers, Comparative evaluations of the geodetic accuracy and cartographic potential of Landsat-4/-5 TM image data, *Photogram. Eng. Remote Sensing*, *51*(9), 1249–1262, 1985.

Westmoreland, S., and D. A. Stow, Category identification of changed land-use polygons in an integrated image processing/geographic information system, *Photogramm. Eng. Remote Sensing*, *58*, 1593–1599, 1992.

Wood, E. F., D.-S. Lin, M. Mancini, D. Thongs, P. A. Troch, T. J. Jackson, J. S. Famiglietti, and E. T. Engman, Intercomparisons between passive and active microwave remote sensing, and hydrological modeling for soil moisture, *Adv. Space Res.*, *13*(5), 167–176, 1993.

CHAPTER 36

FLOODS

STEVEN JENNINGS AND EVE GRUNTFEST

1 INTRODUCTION

The interface between humans and hydrologic features across Earth's surface has helped shape human culture. From the earliest agricultural, complex societies established along some of the great rivers of the world to the bustling seaports of today, humans have gained from the myriad advantages of living in proximity to water. Fertile soil, ease of transportation, and availability of resources (both materials and energy) have allowed for the development of complex material and intellectual cultures. The relationship between water and humans also brings a great deal of risk. Flooding is one of these risks. The impact of floods on humans has been evident from *Genesis* to tonight's evening news. Early Mesopotamian maps may have been drawn to facilitate the reestablishment of property lines after flooding. While the impacts of flooding on humans have been positive in the case of fertile floodplains that support much of the world's agricultural productivity, there is the potential for a great deal of negative impact (Brown, 1984; Clark et al., 1985). Losses of life and property have focused the efforts of scientists, engineers, and government agencies on the prediction, control, and mitigation of floods and flood damage.

In spite of efforts to deal with flooding problems, monetary losses continue to rise at an alarming rate. In Venezuela in December 1999, two weeks of heavy rain resulted on December 15th in flash floods laden with soil, vegetation, and debris. Damages of US$3.2 billion, or 3.3% of the country's gross domestic product were reported. At least 20,000 people were killed. Generally, the number of lives lost due to flooding remains high. However, improvements in flood warnings particularly for major large-scale storms such as cyclones and typhoons have had dramatic effects. The severe 1991 cyclone in Bangladesh resulted in 140,000 dead and property losses

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

of US$2.0 billion. A cyclone of similar intensity in 1993 resulted only in the loss of 126 lives. The early warnings and cyclone shelters accounted for the major improvement (*www.ndndr.org*). China has also witnessed a reduction in the number of lives lost to floods. While 3000 people died in 1998 floods, the 1998 floods were as great as those of 1931 and 1954 where the loss of lives was 145,000 and 33,000, respectively (*www.ndndr.org*). In October 1998 hurricane Mitch was the worst in the eastern Caribbean since 1780 when a hurricane killed 22,000 people. The death toll from Mitch is reported as 11,000. More than 3 million people were left homeless or were severely affected (*www.ncdc.noaa.gov/ol/reports/mitch/mitch.html*).

Floods take a variety of forms with the interplay of several factors leading to the inundation of normally dry land. Wohl (2000) identifies four primary challenges in reducing escalating flood damages. These are (1) estimating flood magnitude for a given recurrence interval, (2) accurately forecasting floods based on rapidly evolving weather conditions, (3) effectively operating flood-warning and evacuation procedures, and (4) establishing and enforcing land-zoning regulations. This chapter first discusses the contexts and causes of flooding, the first two points addressed by Wohl (2000). The second topic is the complex human responses to floods, points 3 and 4 of Wohl (2000). Many of these topics are illustrated with examples of floods.

2 DEFINITION OF FLOODS

Streams are linear water features that flow under the impetus of gravity. The amount of water contained in a stream is usually regulated by contributions of groundwater and surface runoff to the stream channel (Zaslavsky and Sinai, 1981; Knighton, 1998). Much of the time water in a stream flows within the confines of its channel. When inputs of water increase sufficiently, stream discharge leaves the stream channel and covers all or parts of the adjacent floodplain. Since the floodplain surface is usually a virtually flat surface and near the elevation of the stream channel, water can easily spread over the floodplain once water exceeds the elevation of the stream's banks. Most floods develop over a period of days or months as discharge increases gradually (Hirschboeck, 1987, 1988). Flash floods by contrast occur suddenly with little warning and are of short duration. Semiarid and arid areas are likely to experience flash floods (Reid and Frostick, 1987; Hassan, 1990). Flooding is not always associated directly with stream channels. Flooding occurs any time when water covers a surface that is normally not under water. Flooding can occur in coastal areas, low lying areas with poor drainage, or locations with inadequate urban drainage systems.

3 FACTORS THAT LEAD TO FLOODING

Floods have a multitude of causes. Some causes are related to what would be considered natural processes that would occur whether humans are present or not. Many causes have been affected by human activities. In some cases the severity of

floods and the types of damage are a direct result of agriculture, urbanization, and the areas selected for development. In all cases flooding is related to increased discharge in stream channels.

Saturated Soil

Much of Earth's surface is covered by a weathered cover of regolith. Whether forming a true soil with well-developed horizons or a weakly developed detrital cover, the regolith is composed of a mix of mineral particles, organic fragments, and pore space. Commonly, much of the pore space is filled with air and to a lesser extent water. When large amounts of precipitation are received in a region, the pore space fills with water as the input of water from precipitation exceeds the output of water from the soil column to the water table. Decreases in infiltration lead to increases in runoff. The lag time between the precipitation event and the arrival of water to stream channels decreases significantly when soil saturation occurs. As a result, peak discharge increases significantly and the likelihood of overbank flow is high (Smith and Ward, 1998). Spatially, soil saturation may occur over large-scale basins, which leads to flooding in large areas. The peak discharge flows downstream and becomes concentrated in higher order streams causing flooding. In many cases saturation follows a period of high amounts of precipitation over a prolonged time period, possibly weeks or months (Wolman and Gerson, 1978; Ward and Robinson, 1990).

Basin Characteristics

Surface characteristics influence infiltration and runoff rates (Roberts, 1989; Kuhnle et al., 1996). Impervious surfaces such as exposed bedrock or a paved road accelerate surface runoff, thus decreasing lag time between the precipitation event and entrance of water into a nearby channel. Urbanized areas, therefore, with large percentages of impervious surface such as roofs, streets, and parking lots coupled with an engineered drainage system designed to move water quickly to stream channels greatly increase the chances that some flooding will occur after a significant precipitation event (Wolman, 1967; Hammer, 1972; Roberts, 1989; Newson, 1992). Conversely, rural areas with large areas of soil, natural vegetation, and the potential for a faster infiltration rate are less likely to have significant flooding resulting from a single precipitation event. Removal of as much as half the forest cover and a decrease of marsh land along the Yangtze River in China has led to increased flooding. Half a billion people, or 45% of China's total population, reside on the banks or floodplains of the Yangtze and the area produces about 42% of China's gross domestic product. In 1998, 79.6 million people in three Chinese provinces were affected by repeat flooding on the Yangtze. The floods killed more than 3000 people. Fourteen million people were evacuated and 21 million were made homeless (*Weather.ou.edu/spark/AMON/v2_n3/News/DR_980819China12.html*).

Topography

Topography will influence the rate at which precipitation will be incorporated as stream discharge (Patton, 1988). Steep, rocky canyon walls have low infiltration rates as well as a great deal of potential gravitational energy that leads to the concentration of discharge during a short period of time (Strahler, 1964). Alluvial plains usually have a much longer lag time between a precipitation event and the introduction of runoff water into a stream channel. When land cover on steeper slopes is affected by perturbations such as wild fire or building-related oversteepening of slopes, the likelihood of mass movement events is greatly increased. These events are usually related to unstable regolith on steep slopes, which is susceptible to failure when sufficient precipitation is received. For example, see Figure 1.

High Amounts of Precipitation

Flooding is created by the delivery of larger than normal amounts of runoff into stream channels (Smith and Ward, 1998, p. 67). Periods of above-average precipitation lead to floods. In some cases seasonal variability leads to great fluctuations in

Figure 1 (see color insert) Quebrada San Julián upstream of Caraballeda showing evidence of recent debris flows and flash floods. Note the high slope angles, large numbers of debris flow scars, and abundance of new alluvium and colluvium in the channel bed and fan surface. See ftp site for color image.

stream discharge. Wet–dry subtropical or monsoonal climates with distinctive seasons of precipitation lead to fluctuations from dry stream channels to potential flooding events. These cyclical events are related to large-scale atmospheric circulation patterns that operate through an annual or longer period. In the midlatitudes, the annual migration of subtropical high pressures and the polar front lead to distinct precipitation patterns. In the tropics, monsoonal flow can lead to large precipitation events (Milne, 1986). On longer time scales El Niño and La Niña events are persistent over several years and can lead to wet or dry conditions over large areas of Earth's surface from the Equator to the midlatitudes (Waylen and Caviedes, 1987; Pearce, 1988; Ely, et al., 1994).

Extended Wet Periods

In many cases flooding is caused by the reception of precipitation over an extended time period, on the order of weeks to months, that leads to the saturation of soils in a large-scale region (Rodda, 1970b; Smith and Ward, 1998). This saturation leads to increased runoff at a time when streams are at capacity (Ward and Robinson, 1990). Additional water introduced to stream channels cannot be conveyed in the channel but is spread across the floodplain. Wet periods are related to synoptic conditions such as the position of the polar front that delivers cyclonic storms in quick succession. Poleward migration of subtropical air masses over continental areas such as the Mississippi River Basin help to supply large amounts of water to be precipitated by frontal activity. For example, see Figure 2. In some locations rainfall may fall on snow-covered or frozen ground (Thomas and Lamke, 1962). These waters are unavailable to the hydrologic cycle as long as they remain in a solid form. In the case of the former, rainfall may accelerate the introduction of water into the stream network as snowmelt augments the precipitation already being received (Kattelman, 1990; Naef and Bezzola, 1990; Caine, 1995). The latter will greatly decrease the infiltration capacity of the soil causing most of the precipitation to quickly enter the stream network (Horton, 1933).

Decaying Tropical Cyclones

Some of the largest precipitation amounts received as the result of a single meteorological event have been associated with the movement of tropical cyclones (e.g., hurricanes, cyclones, and typhoons) poleward and over continents. These powerful cyclonic storms carry large amounts of warm moist air over land surfaces. While wind speeds associated with these storms decrease quickly after landfall, these decaying storms are capable of delivering precipitation over wide areas during a relatively short period of time, on the order of days to weeks. In some cases cyclonic storms associated with the polar front may exacerbate conditions by introducing a lifting mechanism that leads to increased condensation and precipitation. The relatively low-lying coastal plain of eastern North America is especially susceptible to damage from these types of storms (Bailey and Patterson, 1975; Hirschboeck, 1988). For example, see Figure 3. In 1998 hurricane Mitch produced as much as 50 to 75

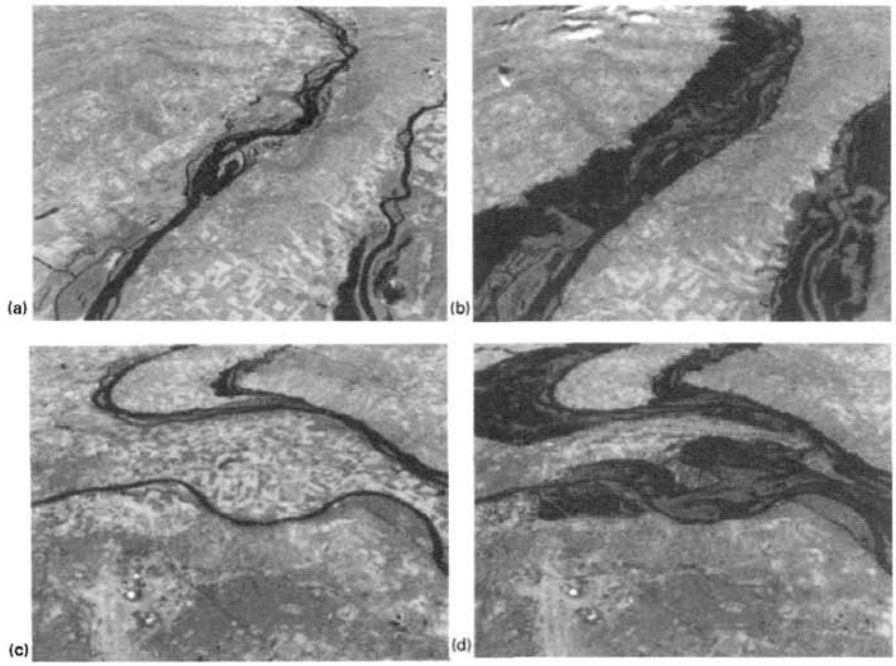


Figure 2 (see color insert) (4 panels): These scenes show various sections of the Mississippi River near St. Louis before and just after the 1993 floods, which peaked in late July/early August. The images show the area as seen by the LandSat Thematic Mapper (TM) instrument. The short-wave infrared (TM band 5), infrared (TM band 4), and visible green (TM band 2) channels are displayed in the images as red, green, and blue, respectively. In this combination, barren and/or recently cultivated land appears red to pink, vegetation appears green, water is dark blue, and artificial structures of concrete and asphalt appear dark gray or black. Reddish areas in the scenes during the flood show where water had started to recede, leaving barren land. See ftp site for color image.

inches of precipitation in some areas of Central America. At least 11,000 deaths were associated with hurricane Mitch and more than 3 million people were left homeless or were severely affected (*www.ncdc.noaa.gov/ol/reports/mitch/mitch.html*).

Intense Thunderstorms

Thunderstorms are usually intense, short-lived storms that produce high winds, hail, and heavy rainfall. These storms can be caused by convection in moist tropical air masses over continental surfaces or fast-moving cold fronts that displace those moist air masses (Hirschboeck, 1987). When these storms develop over mountainous areas where the precipitation is concentrated by the topography the potential for large, catastrophic floods is great (Hall, 1981). For example, see Figure 4. The eastern slope of the Rocky Mountains and the southwestern deserts of North America are

Figure 3 Water and sand washed inland to make travel difficult in North Topsail Island, North Carolina, after hurricane Fran. See ftp site for color image.

common locations for the development of thunderstorms. As moist air encounters higher elevations in these locations, it is forced to rise. Unstable atmospheric conditions are created as mountain slopes heat and in turn heat the atmosphere. Adiabatic cooling causes condensation and the development of large cumulonimbus clouds that can reach the upper altitudes of the troposphere. Sometimes there is little movement associated with a thunderstorm or thunderstorm complex, with respect to the ground; heavy precipitation concentrated in a small geographical area can have catastrophic results.

Quick Snowmelt

The storage of water in the form of snow temporarily removes that water from the hydrologic cycle. In many cases this sequestration of water is short term. Snow accumulates during winter especially at higher elevations and latitudes. With the onset of warmer spring and summer conditions, snowmelt supplies water to streams. A typical early warming may mean that snowmelt may be accelerated with large amounts of runoff entering stream channels. Mountain ranges in mid-latitude coastal regions such as the Coast Ranges and Sierra Nevada of California receive a signifi-

Figure 4 Arizona flash flood, Wenden, Arizona. This community was flooded twice in late October 2000 when waters from Centennial Wash swept into the town. (Photo courtesy of U.S. Small Business Administration.) See ftp site for color image.

cant portion of their annual precipitation in the form of snow. It is possible for warm early spring rains to fall on the snowpack, causing much faster runoff than normal (Bolt et al., 1975; Church, 1988). Another source of snowmelt is the subsurface introduction of heat from volcanic activity. Large volcanoes can be high enough to support permanent snow and ice cover. High temperatures associated with volcanic activity lead to the instantaneous melting of snow and ice. The melt water is commonly mixed with pyroclastic debris to form lahars (Smith, 1996).

Failure of Flood Control Structures

A variety of humanly constructed structures are used in an effort to limit the extent and severity of flooding (Gregory, 1995). Dams and levees are common flood control structures designed to contain water within designated areas (Brookes, 1985, 1988). These structures can fail because of construction errors, poor design, and overtopping by water (Biswas and Chatterjee, 1971; Costa, 1988). Flood control structures can fail because of the failure of a key component. For example, a spillway that erodes away has the potential to lead to the catastrophic failure of the entire dam as the water cuts downward. Sound structures may fail when the water retained by the structure exceeds the height of the structure. Large precipitation events or the displacement of water in a reservoir have the potential to send water flowing over the flood control structure (Kiersch, 1964). This may lead to the failure of the structure

through erosion. Flooding may be exacerbated by these structures since a feature such as a levee tends to raise the stream level well above the floodplain. When a levee fails, a large amount of fluvial energy is concentrated through that break and a great deal of damage can occur near the break.

Cyclonic Storm Created Surges

In low-lying coastal locations a temporary increase in sea level associated with the approach and landfall of storms with significantly high winds and low central pressure can cause significant damage. Sea level rises in response to low pressure as it passes over ocean surfaces. Additionally, the upper portion of the water column is pushed into waves by the high winds. Storm surges can be more than 5 m above the normal high tide (Rappaport, 1994). In some areas such as bays coupled with low-lying deltas, like the Bay of Bengal and the mouth of the Ganges, where storm energy is concentrated, storm surges can reach high levels causing significant flooding (Frank and Husain, 1971; Murty and Neralla, 1992). Barrier islands are also susceptible to flooding by storm surges. Development on barrier islands along the southeastern coast of North America has led to rising property damage related to storms.

Mass Movement Events

A variety of mass movement events, while strictly not fluvial events, behave in a similar way to floods (Carson, 1976). The gravitationally fueled downhill movement of poorly consolidated regolith results from the introduction of meteoric water that adds weight and decreases hillslope cohesion. These events can do significant damage. Several types of mass movement events are composed of a larger percentage of sediments than a typical stream. Events such as mudflows, or lahars, commonly may approach the viscosity and velocity of streams. Valleys can be filled with fine-grained sediments as the deposits dewater following the initial surge of water and sediment. A variety of factors lead to mass movement events. The removal of plant cover by fire may expose soil surfaces so that infiltration rates may increase and lead to the accumulation of water along failure planes in the regolith. In areas with a subtropical wet–dry climate, such as the Mediterranean climate type, the burning of plant cover during the dry season and a subsequent wet season before the reestablishment of plant cover leads to mass wasting events (Rice et al., 1969; Campbell, 1975).

Human Responses to Flooding

There are no accurate estimates of the population in the world's floodplains. Even in the United States, only broad estimates are available, but the trends to increased vulnerability are clear. In 1955 U.S. floodplains had 10 million occupants. Thirty years later the number doubled to 20 million and by the mid-1990s about 12% of the national population lived in areas of periodic inundation. One sixth of the nation's

floodplains are urbanized, and they contain more than 20,000 communities susceptible to flooding. Half of these communities have been developed since the early 1970s (Burby, 1985; Montz and Gruntfest, 1986; Alexander, 1993).

Many of the people at risk do not understand the potential consequences of the hazards they face. In the United States, flood damages exceed $2 billion annually. Only 20 to 30% of eligible structures are insured against flooding. Federal and state disaster assistance accounts for most of the difference. In the United States, almost two-thirds of the residential flood losses result from events that occur once every 1 to 10 years, even though the 100-year floodplain regulation is standard (Alexander, 1993).

In the United States, floods tend to be repetitive phenomena. From 1972 to 1979, 1900 communities were declared disaster areas by the federal government more than once, 351 were inundated at least three times, 46 at least four times and 4 at least five times. As of 1993, the United States was said to spend $9 billion a year on flood control and $300 million on flood forecasting (Alexander, 1993; Conrad, 1998).

Definitions of Structural and Nonstructural Measures

Adjustments to floods can be broadly classified into structural and nonstructural measures. Nonstructural approaches involve adjustment to human activity to accommodate the flood hazard (White, 1964; James, 1975; White, 1974) whereas structural methods are based on flood abatement or the protection of human settlement and activities against the ravages of inundation.

Structural change involves modification to the built environment to minimize or eliminate flood damage directly or flood channel construction changes. For example, see Figure 5. Structural measures are expensive. They may give the illusion of security but the record shows otherwise (Alexander, 1993). The security can be temporary. A flood can occur that is bigger than the design of the channel or levee, and changing priorities in flood control projects that require higher reservoir levels for recreation or water supply can diminish the efficacy of structural measures (Williams, 1998).

The failure of structural flood control works poses a significant threat to the lives of the people who live downstream from a massive structural project such as a dam. More than 2000 people died in 1969 in Italy when the Vaiont Dam collapsed (Blaikie et al., 1994). Because of stringent engineering standards and a system of inspections, the United States has seen few major failures. However, many structures are at the end of their design lives of 50, 75, or 100 years.

Structural flood control is still the dominant idea in many parts of the world. Following the 1927 Mississippi River floods, when river levees collapsed and 200 people died, 700,000 were displaced, and more than 135,000 buildings were damaged (Moore and Moore, 1989), the Army Corps of Engineers did not abandon its dream of controlling all floods. Rather, it proposed building large dams upstream to reduce flood peaks to the capacity of the floodway between the levees (Williams, 1998).

Until the 1970s, most flood loss reduction efforts involved structural solutions. Although nonstructural measures were discussed as alternatives, they were rarely implemented. The shift from mostly structural to mixed structural/nonstructural

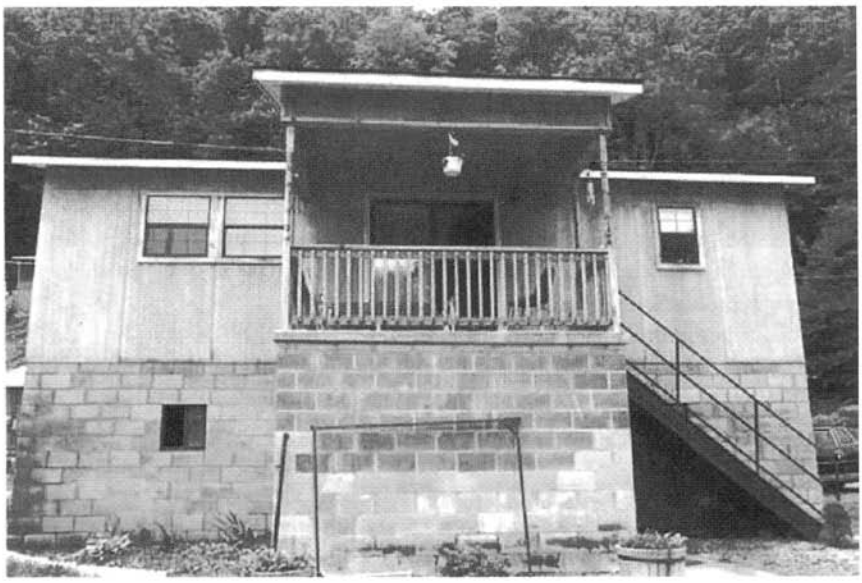


Figure 5 Elevated home in West Virginia is a mitigation success story. Risk is greatly reduced to homes elevated before a flood. See ftp site for color image.

measures began in the 1970s and continues today. The mix of adjustments varies for each situation. In Europe almost all measures that are taken have elements of combined structural and nonstructural measures. There has also been a move to be antistructural. Some dikes are being removed in favor of nonstructural or more environmentally sensitive techniques (Smith and Ward, 1998).

Nonstructural measures include floodproofing, land-use planning, soil bioengineering, warning systems, preflood mitigation efforts, and insurance. The simplest nonstructural measure is to accept the loss. Another nonstructural measure is to provide postflood relief. Protection of floodplain residents and users, and the supply of relief when they suffer damage, are forms of hidden subsidy (Alexander, 1993). This category includes aid provided by the Red Cross, voluntary organizations, and governmental agencies.

Nonstructural measures include flood insurance and land-use management, acquisition and relocation, floodproofing, preflood mitigation preparedness, outdoor warning systems, and soil bioengineering,

4 DISCUSSION OF NONSTRUCTURAL MEASURES

Flood Insurance, Floodplain Mapping, and Land-Use Ordinances

In 1968 the U.S. National Flood Insurance Program (NFIP) was launched. It made affordable insurance available to residents in flood-prone areas. In 1999 more than

18,000 communities belonged to the program. Participating local governments require developers to meet minimum standards designed to avoid damages that might be inflicted by a catastrophic 100-year flood. The program also requires property owners to purchase flood insurance to receive a federally insured mortgage (Myers, 1996). Flood insurance is a means for placing some of the burden of losses onto the people who take (or make) the risk, namely the floodplain users and residents (Alexander, 1993). Communities can participate in a Community Rating System, established by the Federal Emergency Management Agency (FEMA), that allows them to show innovative strategies to reduce flood losses in return for lower insurance premiums for floodplain residents.

Before a community can participate in the flood insurance program, the flood hazard must be recognized, assessed, and mapped. These assessments include flood history, cost and types of past flood damages, maps of the limits of the 100-year flood (or other designated flood) on a topographic map, compilations of profiles and cross sections of the river to show the levels of past floods, and compilations of flood frequency curves and locally representative hydrographs.

FEMA works with the state and community governments to identify their flood hazard areas and publishes a Flood Hazard Boundary Map of those areas. When a community joins the NFIP, it must require permits for all construction or other development in these areas and ensure that the construction materials and methods used will minimize flood damage. However, there is not careful monitoring to be sure that reducing flood hazard in a particular area does not increase flood potential elsewhere. Often, the problems are just shifted to different locales. In return the federal government makes subsidized flood insurance available to those whose structures were in the flood hazard area prior to issuance of the flood maps. All others are eligible for flood insurance at actuarial rates. FEMA issues a Flood Insurance Rate Map after the Flood Insurance Study of risk zones and elevations has been prepared (*http://floodplain.org/Jan32.htm*).

Acquisition and Relocation

The most effective measure to reduce losses is to keep the floodplains free of development. However, in many river valleys in the world, it is too late for that option. One of the most promising strategies for reducing flood losses is the public acquisition of developed land susceptible to flooding (Conrad, 1998; *www.fema.gov/mit/homsups.htm*). The authorization for U.S. federal cost sharing for relocation is more than 30 years old. However, only recently have communities, tired by chronic flooding, taken advantage of funding packages and relocated. In one case, the entire town of Valmeyer, Illinois, was relocated. The town had a long history of floods. In 1943, 1944, and 1947 unusually high levels of the Mississippi caused flooding in the nearby bottomlands affecting Valmeyer. After the 1947 floods, the U.S. Army Corps of Engineers raised the levees protecting the reach of the floodplain to 47 ft. On August 1, 1993, the flood overtopped the levees inundating Valmeyer, prompting its ultimate relocation. Since 1993 nearly 20,000 properties

in 36 states and one territory have been bought out and over 25,000 families have moved from floodplains (*http://www.nwf.org/nwf/pubs/higherground/intro.html*).

Floodproofing

Floodproofing is a range of adjustments aimed at reducing flood damages to a structure or to the contents of buildings. There are three categories: (1) raising or moving the structure; (2) constructing barriers to stop floodwater from entering a building; and (3) wet flood proofing (U.S. Army Corps of Engineers, 1997).

Detection and Response Warning Systems

New technological advances in stream and rain gage networks and the increased regional floodplain management efforts have led to the adoption of thousands of local flood-warning systems. Many are simple detection systems and do not provide any mechanism for alerting the population at risk. In the United States until the 1990s warning or detection systems were planned and administered primarily at the local level.

Since then, the federal government including the Bureau of Reclamation, the U.S. Army Corps of Engineers, the National Oceanic and Atmospheric Administration, and the Federal Emergency Management Agency have actively participated in the installation and maintenance of detection and warning systems. Many systems are still managed by regional or local entities, but the percentage of federal dollars has increased substantially. Standards have also been established to help make the systems more compatible across regions (U.S. Department of Commerce, 1997).

An automated integrated network of stream and rain gages is being used in more than 1000 communities in the United States to help provide lead time for floods. Most of the systems are developed through collaborative efforts of many agencies. These ALERT systems (automated local evaluation in real time) have performed many functions other than flood warning, including helping in water supply decision making, fire weather forecasting, pollution monitoring, and providing data for river recreationists (Gruntfest and Huber, 1991). The availability of real-time data on the Internet also has increased interest in these monitoring systems (Gruntfest and Weber, 1998). The State of Arizona is developing a network for flood warning throughout the state. More than 30 agencies and communities are working together on the comprehensive ALERT system (*http://www.alertsystems.org/saas/*).

Warning systems may be nothing more than "cheap payoffs of the raingods." Too often communities install rain gage/stream gage monitoring systems without a plan for getting the warning message disseminated. A warning system is only necessary once poor land-use decisions have been made, allowing people to settle in harm's way. Many of the systems being built are not being adequately maintained to be reliable (Gruntfest and Huber, 1991; Parker and Fordham, 1996). Public education encouraging people to heed environmental cues is also being used. It is particularly difficult to provide adequate lead times for flash floods. Some communities do have

drills to test the reliability and completeness of their systems to be sure the systems will operate when the conditions warrant.

As of 2001 a combination of factors increase the likelihood that automated detection systems may become more popular and more valuable. More powerful, less expensive computers, and World Wide Web access provide opportunities for inexpensive real-time weather data. While real-time stream and rain gage networks may be originally installed for flash flood forecasting, many agencies and users find the data useful for alternative purposes.

Soil Bioengineering

Anchored plantings along stream banks serve as the basis for this technique. Soil bioengineering and biotechnical engineering are cost-effective and environmentally compatible ways to protect slopes against surficial erosion and shallow mass movement. These approaches provide alternatives to structural channel "improvements." They raise questions about the notion of why engineers ever considered that concrete-lined channels should be considered "improved" (Gray and Sotir, 1996). Generally, bioengineering solutions must also include a strategy to carry floodwaters away.

The bioengineering technique is gaining support throughout the United States and Europe. It is less expensive to install and less expensive to maintain as well. The broader adoption of soil bioengineering may radically alter floodplain management.

Combined Structural and Nonstructural Measures to Reduce Flood Losses

From the first attempts to reduce flood losses in the United States, structural measures were preferred for three main reasons: (1) their benefits appeared to be relatively easy to measure, (2) they did not require extensive and politically controversial land-use planning, and, (3) the federal cost-sharing agreements encouraged communities to select the most expensive engineering projects. These reasons were supported by a faith in the technology of structural measures to protect people and property from floods.

The record now shows that in spite of massive expenditures, flood losses have continued to rise. Since the 1960s, especially in the United States, there has been a call for a shift from primarily structural measures to control floods to nonstructural measures (Galloway, 1994; Larson, 1996; Williams, 1998). Land-use control is one of the most effective ways of reducing flood hazards. Statutes, ordinances, regulations, and compulsory purchases can be employed and relocation can be subsidized. A floodway left undeveloped through the city can become beautiful public open space.

5 CONCLUSION

Floods are generally caused by the combination of large amounts of precipitation and basin topography. For example, saturation of soils caused by large amounts of

precipitation can lead to flooding. Urbanization of a drainage basin increases the amount of runoff reaching a channel and decreases the lag time between a precipitation event and peak flow. A variety of weather events lead to flooding, including extended wet periods, decaying tropical cyclones, intense thunderstorms, and quick snowmelt. In some cases humanly constructed structures designed to prevent flooding collapse, causing flooding or accentuating flooding. In low-lying coastal areas storm surges may cause significant flooding. Mass movement events are similar to flooding, although the proportion of sediments to water is larger than an alluvial flood with the outcome just as disastrous.

Humans respond to flooding in a variety of ways. Broadly defined these fall into two categories, structural and nonstructural measures. Structural measures include dams and dikes. Through time the efficacy of structural features has been questioned, and there has been a shift from purely structural approaches to controlling floods to a mix of structural and nonstructural flood mitigation strategies. Nonstructural measures include flood insurance, floodplain mapping, and land-use ordinances, acquisition and relocation, floodproofing, detection and response warning systems, soil bioengineering, and combined structural and nonstructural measures to reduce flood losses. Some progress is being made in addressing the hazards associated with flooding. The reduction of flood impacts continues at great expense, but vulnerability will continue to rise as long as more people build in floodplains, increasing the risk of catastrophic floods. Even the best warnings will not eliminate the risks increasingly being taken around the globe.

References

Alexander, D., *Natural Disasters*, Chapman and Hall, New York, 1993, 632 p.

Bailey, J. F. and J. L. Patterson, *Hurricane Agnes Rainfall and Floods, June–July 1972*, USGS Professional Paper 924.

Biswas, A. K. and S. Chatterjee, Dam disasters: an assessment, *Engineering Journal, 54*, 3–8, 1971.

Blaikie P., T. Cannon, I. Davis, and B. Wisner, *At Risk Natural Hazards, Peoples Vulnerability and Disasters*, Routledge, New York, 1994, 282 p.

Bolt, B. A., W. L. Horn, G. A. Macdonald, and R. F. Scott, *Geological Hazards*, Verlag, Berlin, 1975.

Brookes, A., River channelization: traditional engineering methods, physical consequences and alternative practices, *Progress in Physical Geography, 9*, 44–73, 1985.

Brookes, A., *Channelized Rivers: Perspectives for Environmental Management*, John Wiley & Sons, Chichester, 1988.

Brown, L. R., Conserving soils, in L. R. Brown (Ed.), *State of the World*, Norton, New York, 1984, pp. 53–75.

Burby, R. J., *Flood Plain Land Use Management: A National Assessment*, Westview Press, Boulder Colorado, 1985.

Caine, N., Snowpack influences on geomorphic processes in Green Lakes valley, Colorado Front Range, *Geographical Journal, 161*, 55–68, 1995.

Campbell, R. H., Soil slips, debris flows and rainstorms in the Santa Monica Mountains and vicinity, southern California, *US Geological Survey, Professional Paper 851*, 1975.

Carson, M. A., Mass-wasting, slope development and climate, in E. Derbyshire (Ed.), *Geomorphology and Climate*, John Wiley & Sons, Chichester, 1976, pp. 101–136.

Church, M., Floods in cold climates, in V. R. Baker, R. C. Kochel, and P. C. Patton (Eds.), *Flood Geomorphology*, John Wiley & Sons, New York, 1988, pp. 205–229.

Clark, E. H., J. A. Haverkamp, and W. Chapman, *Eroding Soils: The Off-Farm Impacts*, The Conservation Foundation, Washington, DC, 1985.

Conrad, D., *Higher Ground: Voluntary Property Buyouts in the Nation's Floodplains, A Common Ground Solution Serving People at Risk, Taxpayers and the Environment*, National Wildlife Federation, Washington, DC, 1998.

Costa, John E., Floods from Dam Failures, in V. R. Baker, R. C. Kochel, and P. C. Patton (Eds.), *Flood Geomorphology*, John Wiley & Sons, New York, 1988, pp. 439–463.

Ely, L. L., Y. Enzel, and D. R. Cayan, Anomalous North Pacific atmospheric circulation and large winter floods in the southwestern United States, *Journal of Climate*, *7*, 977–987, 1994.

Frank, N. L. and S. A. Husain, The deadliest tropical cyclone in history? *Bulletin of American Meteorological Society*, *52*(6), 438–444, 1971.

Galloway, G., *A Blueprint for Change, Sharing the Challenge: Floodplain Management into the 21st Century*, report of the Interagency Floodplain Management Review Committee to the Administration Floodplain Management Task Force, Washington, DC, June 1994.

Gray, D. H. and R. B. Sotir, *Biotechnical and Soil Bioengineering Slope Stabilization: A Practical Guide for Erosion Control*, Wiley, New York, 1996, 378 p.

Gregory, K. J., Human activity and paleohydrology, in K. H. Gregory, L. Starkel, and V. R. Baker (Eds.), *Global Continental Palaeohydrology*, Wiley-Interscience, Chichester, 1995, pp. 151–172.

Gruntfest, E. and C. J. Huber, Toward a comprehensive national assessment of flash flooding in the United States, *Episodes*, *14*(1), 26–34, 1991.

Gruntfest, E. and M. Weber, Internet and emergency management prospects for the future, *International Journal of Mass Emergencies and Disasters*, *16*(1), 1998.

Hall, A. J., *Flash Flood Forecasting*, Operational Hydrology Report No. 18, WMO, Geneva, 1981.

Hammer, T. R., Stream channel enlargement due to urbanization, *Water Resources Research*, *8*, 1530–1540, 1972.

Hassan, M. A., Observations of desert flood bores, *Earth Surface Processes and Landforms*, *15*, 481–485, 1990.

Hirschboeck, K. K., Catastrophic flooding and atmospheric circulation anomalies, in L. Mayer and D. Nash (Eds.), *Catastrophic Flooding*, Unwin Hyman, London, 1987, pp. 23–56.

Hirschboeck, K. K., Flood hydroclimatology, in V. R. Baker, R. C. Kochel, and P. C. Patton (Eds.), *Flood Geomorphology*, John Wiley & Sons, New York, 1988, pp. 27–49.

Horton, R. E., The role of infiltration in the hydrologic cycle. *Transactions of the American Geophysical Union*, *14*, 446–460, 1933.

Kattelman, R., Floods in the high Sierra Nevada, California, USA, in R. O. Sinniger and M. Monbaron (Eds.), *Hydrology in Mountainous Regions II. Artificial Reservoirs, Water and Slopes*, IAHS Publ. No. 205, 1990, pp. 311–317.

Kiersch, G. A., The Vaiont River disaster, *Civil Engineering*, *34*, 32–39, 1964.

Knighton, David, *Fluvial Forms and Processes: A New Perspective*, Arnold, London, 1998.

Kuhnle, R. A., R. L. Binger, G. R. Foster, and E. H. Grissinger, Effect of land use changes on sediment transport in Goodwin Creek, *Water Resources Research*, *32*, 3189–3196, 1996.

Larson, L., Lessons drawn from the 1993 flood, *Forum for Applied Research and Public Policy*, Fall, No. 3, 1996, pp. 102–104, 1996.

Milne, A., *Flood Shock*, Sutton, Gloucester, 1986.

Montz, B. E. and E. Gruntfest, Changes in American urban floodplain occupancy since 1958: the experiences of nine cities, *Applied Geography*, *6*, 325–338, 1986.

Moore, J. W. and D. P. Moore, *The Army Corps of Engineers and the Evolution of Federal Floodplain Management Policy*, Institute of Behavioral Science, Boulder, CO, special publication No. 30, 1989.

Murty, T. S. and V. R. Neralla, On recurvature of tropical cyclones and the storm surge problem in Bangladesh, *Natural Hazards*, *6*(3), 275–279, 1992.

Myers, M. F., Midwest floods channel reforms, *Forum for Applied Research and Public Policy*, *11*, Fall, No. 3, 1996, pp. 88–97.

Naef, F. and G. R. Bezzola, Hydrology and morphological consequences of the 1987 flood event in the upper Reuss valley, in R. O. Sinniger and M. Monbaron (Eds.), *Hydrology in Mountainous Regions II. Artificial Reservoirs, Water and Slopes*, IAHS Publ. No. 205, 1990, pp. 339–346.

Newson, M., *Land, Water and Development: River Basin Systems and their Sustainable Management*, Routledge, London, 1992.

Parker, D. and M. Fordham, Evaluation of flood forecasting, warning and response systems in the European Union, *Water Resources Management*, *10*(4), 1996, pp. 279–302.

Patton, Peter C., Drainage basin morphometry and floods, in V. R. Baker, R. C. Kochel, and P. C. Patton (Eds.), *Flood Geomorphology*, John Wiley & Sons, New York, 1988, pp. 51–64.

Pearce, F., Cool oceans caused floods in Bangladesh and Sudan, *New Scientist*, *8*, 31, 1988.

Rappaport, E. N., Hurricane Andrew, *Weather*, *49*(2), 51–60, 1994.

Reid, I. and L. E. Frostick, Flow dynamics and suspended sediment properties in arid zone flash floods, *Hydrologic Processes*, *1*, 239–253, 1987.

Rice, R. M., E. S. Corbett, and R. G. Bailey, Soil slips related vegetation, topography and soil in southern California, *Water Resources Research*, *5*, 647–659, 1969.

Roberts, C. R., Flood frequency and urban-induced channel change: Some British examples, in K. Beven and P. Carling (Eds.), *Floods: Hydrological, Sedimentological and Geomorphological Implications*, John Wiley & Sons, Chichester, 1989, pp. 57–82.

Rodda, J. C., Rainfall excesses in the United Kingdom, *Transactions of the Institute of British Geography*, *49*, 49–60, 1970.

Smith, K. and R. C. Ward, *Floods: Physical Processes and Human Impacts*, John Wiley, New York, 1998.

Smith, K., *Environmental Hazards: Assessing Risk and Reducing Disaster*, 2nd ed., Routledge, London, 1996.

Strahler, Alan N., Quantitative geomorphology of drainage basins and channel networks, in V. T. Chow (Ed.), *Handbook of Applied Hydrology*, McGraw-Hill, New York, 1964, pp. 4-40–4-74.

Thomas, C. A. and R. D. Lamke, Floods of February 1962 in southern Idaho and northeastern Nevada, *US Geological Survey Circular*, *467*, 30, 1962.

U.S. Army Corps of Engineers, *Flood Proofing Techniques, Programs and References*, Available from Corps at CECW-PF, 20 Massachusetts Avenue, NW, Washington, DC, 1997, 26 pp.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, Office of Hydrology, *Automated Local Flood Warning Systems Handbook*, Weather Service Hydrology Handbook No. 2, February, Silver Spring, MD, 1997, unpaged.

Ward, R. C. and M. Robinson, *Principles of Hydrology*, 3rd ed, McGraw-Hill, Maidenhead, 1990.

Waylen, P. R. and C. N. Caviedes, El Nino and annual floods in coastal Peru, in L. Mayer and D. Nash (Eds.), *Catastrophic Flooding*, Unwin Hyman, 1987, pp. 57–77.

Williams, P., Inviting trouble downstream, *Civil Engineering*, February, 1998, pp. 50–53.

Wohl, Ellen (Ed.), *Inland Floods*, Cambridge University Press, 2000.

Wolman, M. G., A cycle of sedimentation and erosion in urban river channels, *Geografiska Annaler, 49A*, 385–395, 1967.

Wolman M. G. and R. Gerson, Relative scales of time and effectiveness of climate in watershed geomorphology, *Earth Surf. Proc., 3*, 189–208, 1978.

Zaslavsky, D. and G. Sinai, Surface hydrology: I-Explanation of phenomena; II-Distribution of raindrops; III-Causes of lateral flow; IV-Flow in sloping layered soil; V-In-surface transient flow, *Journal of Hydraulics Division American Society of Civil Engineers, 107(HY1)*, 1–93, 1981.

Selected Web Pages Related to Nonstructural Measures

http://www.alertsystems.org/saas/ALERT User Group

http://FEMA.gov U.S. Federal Emergency Management Agency

http://ceres.ca.ca.gov/ – State of California water resources agency

http://www.usace.army.mil/inet/functions/cw/cwfpms/fpms.htm U.S. Army Corps of Engineers with emphasis on floodplain management activities

http://www.nwf.org/nwf/pubs/higherground/intro.html National Wildlife Federation site for manuscript Higher Ground

http://member.aol.com/damsafety/homepage.htm Association of State Dam Safety Officials

http://www.ci.fort-collins.co.us/csafety/oem/index.htm Comprehensive emergency preparedness homepage from Ft Collins, Colorado an excellent reference.

http://web.uccs.edu/geogenvs/work/Eve/Beyond%20Flood%20Detection%20Final.html.

www.ncdc.noaa.gov/ol/reports/mitch/mitch.html – NOAA Website about Hurricane Mitch

SECTION 3

SOCIETAL IMPACTS

CHAPTER 37

CLIMATE AND SOCIETY

MICHAEL H. GLANTZ

At the turn of the twentieth century, scholars who wrote about the interplay between climate and society did so based on their perceptions of climate as a boundary constraint for the development prospects of a society. Perceptions of climate were used as an excuse to dominate societies in Africa, Asia, and Latin America. As a result, climate–society studies soon became viewed as a colonial ploy to control populations in developing areas in the tropics. Perhaps the most cited book in this regard was written by Ellsworth Huntington, *Climate and Civilization*, published in 1915. In his view, inhabitants of the tropics were destined to accept lower levels of economic and social development because their climate setting was not conducive to lively (i.e., productive) human activity or an aggressive work ethic. According to Huntington, tropical climate was the main culprit causing people in the tropics to be less productive than people in temperate regions. Huntington argued that the temperate climate has an energizing effect on humans. With the growing belief that such an argument was racist in intent, Huntington's work was challenged, and discussion of the various ways in which climate might influence human behavior was stifled for decades, notwithstanding a few notable exceptions. One such exception is entitled *Climate and the Energy of Nations* (Markham, 1944) in which Markham referred to the "air-conditioning revolution," a revolution based on the development and spread of a new technology into the tropical areas. Markham asserted that technology brings islands of temperate-zone climate into the tropics, thereby generating a more aggressive work ethic.

Following the end of World War II and the onset of the Cold War between Soviet-style communism and Western capitalism and democracy, attention of governments turned to Cold War conflicts, avoidance of nuclear war, searches for allies, and decolonization. The major Cold War nations were in a competition to show that *their* approach to economic development was the only way for the newly indepen-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

dent countries to follow. A main stated objective was their intent to assist these countries to become food secure based on the nation's resources. Consideration of climate was making its way back into the discussions of economic development in developing countries. Once again interest was raised with regard to climatic constraints to economic development in tropical countries.

In the 1950s and 1960s, attention focused on decolonization and political development of the newly independent states (e.g., Pye, 1966). In the mid-1970s, a World Bank report about the economic prospects for developing countries—*The Tropics and Economic Development: A Provocative Inquiry into the Poverty of Nations*—hinted at the economic, social, and political problems caused by climate variability from one year to the next. Its author (Kamarck, 1976) noted that recurrent droughts in northeast Brazil are a chronic constraint on the region's economic development prospects. His reference to interannual climate variability was brief and unelaborated. However, climate as a boundary constraint was starting to give way to climate as something that societies might be able to forecast and cope with, at least in its extremes.

In the 1970s, attention focused on how the vagaries of weather exposed hundreds of millions of people to hunger and, depending on the socioeconomic situation in a particular country, to famine as well (e.g., Glantz, 1976; Sen, 1981). Thus, there was a growing number of examples of the notion that climate was not really a boundary constraint to the level of development that a people or culture could expect to attain. This notion began to give way to the belief that variability in climate, from one year to the next or one decade to the next, could be coped with so as to soften the impacts of climate variability and weather extremes on agriculture and livestock and, more generally, on the productivity of the land's surface (e.g., Glantz, 1977; Hare, 1977).

Recall that the 1970s was a disruptive decade with respect to climate: 5 years of drought in the West African Sahel (Glantz, 1976); failure of the Soviet harvest and subsequent large-scale, low-cost grain purchases by the Soviet Union in the early 1970s (Trager, 1975); the global food crisis (Brown and Eckholm, 1974); talk of a possible return to an ice age (e.g., Ponte, 1976; Weather Conspiracy, 1974); the Ethiopian famine (Wolde Mariam, 1988); drought-related coups in sub-Saharan Africa; drought in the wheat-producing Canadian prairie provinces (Glantz, 1977); the first drop in global fish catches since the end of World War II (Brown and Eckholm, 1974), and so forth.

A devastating 5-year drought from 1968 to 1973 in the West African Sahel and its associated death and environmental destruction in the region drew attention to the impacts on household and village responses to prolonged, multiyear droughts. Widespread droughts around the globe in 1972–1973, famines in West Africa and Ethiopia, blamed for the most part on an El Niño event, along with the drop in fish landings, prompted the U.N. Secretary General to convene a series of UN-sponsored world conferences on food (1974), population (1974), human settlements (1976), water (1977), desertification (1977), climate (1979) and technology (1979).

Thus, toward the middle of the 1970s, at least five new major climate-related scientific issues emerged: the effect of chlorofluorocarbons (CFCs) on the ozone layer in the stratosphere, talk of an impending Ice Age suddenly shifted to talk of a

human-induced global warming, acid rain, desertification, and El Niño. Each of these issues raised interest in climate–society interactions to higher levels among researchers in different disciplines, government agencies, economic sectors, the media, and the public. Societies around the globe responded (and continue to respond) in different ways to each of these climate-related issues. For example, desertification is an environmental issue that is of great concern to African countries.

North Americans, however, refused to accept the view that desertification could occur in the U.S. West as a result of mismanagement of the land's surface, while noting that desertification was the plight of poor developing countries in Africa. The term *desertification* first appeared in a report on the destruction of dry forests in central Africa by a French forester (Aubreville, 1949). Since then, the concept of desertification has been expanded to include such land degradation processes as soil erosion, wind deflation, soil salinization, water logging, livestock overgrazing, and soil trampling. While many of these processes were exposed during the prolonged drought in the West African Sahel and then labeled as desertification, it is not difficult to show that similar processes also take place in the U.S. West.

The acid rain issue was addressed in the United States with the implementation by the U.S. Congress of a decade-long national assessment called NAPAP (National Acid Precipitation Assessment Program). Stratospheric ozone depletion was addressed globally in the 1980s with the development of international legal instruments culminating in the Montreal Protocol of 1987 and, later, amendments to it (Benedick, 1998).

It was in the early 1970s, 1972–1973 to be exact, that an El Niño event (defined briefly as an invasion of warm water from the Western Pacific into the central and eastern equatorial Pacific Ocean) attracted global attention. An event in 1982–1983, the biggest in a century until that time, captured the full attention of the scientific community and various governments as a natural phenomenon that spawned hazards around the globe. Such hazards included, but were not limited to, droughts, floods, frosts, fires and food shortages, famine, and disease. Ever since the mid-1970s, research funding of El Niño–related research has been growing along with international interest in the phenomenon and its societal and environmental impacts. The extraordinary El Niño event of 1997–1998 helped to make El Niño and its cold counterpart, La Niña, household words throughout much of the world. Only at the end of the twentieth century did La Niña events become of serious interest to the El Niño research and forecasting communities (Glantz, 2002). This belated interest is even more surprising given the scientific observation that tropical storms and hurricanes in the Atlantic Basin and in the Gulf of Mexico tend to increase in number during La Niña events and drop in number during El Niño events.

Global warming is an environmental issue that arose out of discussions and governmental and scientific concerns about the possibility of a global cooling. It was first suggested in 1896 by Swedish chemist Arrhenius (1896, 1908) that the burning of coal by human societies would add enough extra carbon dioxide into the air to eventually heat up Earth's atmosphere by a few degrees Celsius. This issue was revisited in the 1930s by Callendar (1938), who thought that a human-induced global warming of the atmosphere could stave off the imminent recurrence of an

ice age. The issue was again revisited in the 1950s when global warming was looked at in neutral terms, as an experiment that societies were performing on the chemistry of the atmosphere, for which the outcome is unknown (Revelle and Suess, 1957).

It was not until the mid-1970s that human-induced global warming began to be viewed as an adverse event for future generations of human societies and ecosystems that might not be able to adapt to the rate of warming expected to occur. The cause of the warming was attributed to the increasing amounts of greenhouse gases (CO_2, CFCs, CH_4, NO_x, collectively referred to as GHGs) being emitted into the atmosphere as a result of human activities. Carbon dioxide is a product of the burning of fossil fuels, and its amount in the atmosphere has been rising since the onset of the Industrial Revolution in the late 1700s. Tropical deforestation also contributes carbon dioxide to the atmosphere. Tropical forests have served as sinks for carbon dioxide, pulling it out of the air and storing it. When trees are felled, decompose or burned, the stored carbon is emitted into the air.

Chlorofluorocarbons (CFCs), a greenhouse gas as well as a stratospheric "ozone eater," are man-made chemicals first discovered in the 1920s for use as a refrigerant. Methane resulting from livestock rearing (e.g., cattle, pigs) and from the increasing number of landfills is another greenhouse gas. Nitrous oxides are used by farmers in fertilizers and have been widely applied to agricultural lands around the globe in increasing amounts since the end of World War II. Of these major greenhouse gases, carbon dioxide is seen at the main culprit in the global warming debate.

Current scientific research suggests that the level of climate change that might be expected (at current rates of greenhouse gas emissions) is on the order of 1.5 to 4.5°C by the end of the twenty-first century (IPCC, 1990, 1996, 2001). Concerned with the prospects of a changing global climate, many nations have come together to call for a technical assessment of the state of the science through the Intergovernmental Panel on Climate Change (IPCC).

The degree of warming, however, is dependent on numerous factors: the rate at which GHGs continue to be emitted into the atmosphere, the shift by societies to alternative energy sources, the rate of tropical deforestation, the residence time of GHGs in the atmosphere (several of these gases will remain in the atmosphere for decades to centuries), the development of methods to sequester carbon (i.e., taking it from the atmosphere and binding it in some way in Earth's land surface, vegetation, or oceans), and so forth. Some degree of global warming is inevitable, given the residence time of the GHGs already emitted into the atmosphere. This means that societies around the globe, from local to national, must attempt to ascertain how a warmer global climate regime might affect regional and local climates. Will there be more extreme climate events (such as droughts, floods, frosts, fires) or fewer? These societies must also seriously consider nationally, as well as collectively in cooperation with other countries, the most effective way(s) to
cope with the potentially adverse impacts of some degree of human-induced global warming.

Coping mechanisms for climate change likely to occur decades in the future can be divided into three categories: preventive, mitigative and adaptive measures. *Preventive* measures are designed to prevent the increased buildup of GHGs in

the atmosphere. *Mitigative* measures depend on an improved understanding of how global warming might affect local climates worldwide and are designed to improve societies' ability to respond to changes, some of which can be anticipated with some degree of reliability. *Adaptive* measures are used to refer to society pursuing a "business as usual" strategy, not seeking to control GHG emissions, allowing global warming to occur, and responding to any of the impacts of climate change as they might appear. Today, adaptation now encompasses mitigation as well (Smith et al., 1996).

As we enter a new millennium, it appears that national governments have shifted their concern from global climate change to local and regional climate impacts, and from climate change alone to climate change AND climate variability on various time scales, from the seasons to years to decades. Reinforcing this interest has been the fact that climate-related disasters in the 1990s have been the most costly since records have been kept.

More specifically, the climate-related events of the 1990s merit special attention. Devastating hurricanes, such as Hurricane Andrew (1992), Hurricanes Mitch and Georges (1998), Hurricane Floyd (1999), floods in Europe (1993 and 1995), floods in Bangladesh (1998), devastating floods in China (1998), flash flooding and mud slides in Venezuela (1999), droughts in southern Africa (1991–1992), a prolonged El Niño event (1991–1995), the destructive ice storm in northeastern North America (1998), and a second El Niño of the century (1997–1998), drought and famine in Ethiopia (2000), among many other climate-related problems, have heightened public, media, and policymaking interest in climate to levels never before seen. While several of these events might have been expected to occur, they were for the most part surprising in their timing or intensity.

As a result of this new-found international concern, there has been an increasing number of studies on climate-related impacts and on how societies have been affected by (or coped with) those impacts. There has also been an apparent realization that an improved understanding of climate variability and climate extremes can help societies to better cope with climate change several decades in the future.

Recent hurricanes, ice storms, droughts, floods, and intense El Niño events have proven that all countries, regardless of their level of economic development or type of political system, are subject to the adverse effects of climate variability. Despite programs designed to "droughtproof" or "floodproof" a country, recent studies have shown that industrialized countries are no more immune to the impacts of climate from one year to the next than are developing countries. A major difference, however, is that in the industrialized countries, governments are in a relatively better position (economically) than developing ones to address those impacts and their long-term implications for economic development prospects.

Climate modelers and other climate researchers have come to realize that there are likely to be more surprises with regard to the behavior of the global climate system. While there are aspects of the system that are somewhat predictable, other aspects are not. A climate-related surprise can be defined as a gap between one's expectations about climate's behavior and what the climate system actually does.

Climate-related surprise is not a black-and-white condition. People are hardly ever either totally surprised or never surprised. There are shades of surprise with regard to human responses to the same climate-related event. They can be hardly surprised, mildly surprised, somewhat surprised, very surprised, extremely surprised or totally surprised (NB: each of these examples was taken from the scientific literature). Myers (1995, p. 358) introduced the interesting notion of "semisurprised." Thus, surprise may best be described in "fuzzy" terms with the degree of surprise dependent on several intervening variables such as personal experience, core beliefs, expectations, or knowledge about a phenomenon or about a geographic location.

One could argue that there are knowable as well as unknowable surprises (Streets and Glantz, 2000). Knowable refers to the fact that some climate surprises can be anticipated (Myers, 1995). For example, certain parts of the globe are drought prone. It is known that drought will likely recur. What is not known is exactly when it will take place, how long it will last, how intense it will be, or where its most devastating impacts are likely to occur. El Niño is in this category. While we have now come to expect these events to recur, we do not know when that will happen or what it will be like. The uncertainty then cascades down the "impacts chain," and as we speculate about likely impacts of an El Niño, the degree of uncertainty will increase.

Take, for example, the 1997–1998 El Niño. Even with the best monitoring and observing system in the world focused on minute changes in various aspects of the tropical Pacific Ocean, forecasters and modelers were unable to predict the onset of one of the biggest El Niño events in the past 100 years. Nor were they able to predict the course of development of that event. They were better than in earlier times, however, at predicting some of its impacts on societies in certain parts of the globe, especially those where the influences of changes in the sea surface temperatures in the tropical Pacific are known to be strong.

Societies (and their scientists) are on a learning curve with regard to the various ways that climate variability and climate change might affect climate-related human activities. They must avoid becoming complacent as a result of a belief that they fully understand atmospheric processes or their impacts. They must accept that there will be climate surprises in the future, even if the global climate does not change. They must learn from past experiences on how best to cope with the vagaries of climate (Glantz, 1988).

Many countries now realize that climate-related problems do not stop at international boundaries. There are many transboundary issues that demand regional (if not international) cooperation, given that countries share river basins, inland seas, airsheds, the global atmosphere as well as the onslaught and impacts of extreme meteorological events such as droughts, floods, and tropical storms.

While climate-related anomalies cannot be prevented, societal preparation for, and response to, their adverse impacts can be improved through better knowledge of the direct and indirect ways in which atmospheric processes interact with human activities and ecological processes. The enhancement of such knowledge will lead to better forecasts as well as better computer modeling of the interactions among land, sea, and air. A society forewarned of climate-related hazards is forearmed to cope with those hazards more effectively.

REFERENCES

Arrhenius, S., *Worlds in the Making*, Harper & Brothers, New York, 1908.

Arrhenius, S., On the influence of carbonic acid in the air upon the temperature of the ground, *Philos. Mag.*, 41, 237276, 1896.

Aubreville, A., *Climate, Forests and Desertification in Tropical Africa*, Sociéte d'Editions Géographiques, Maritimes et Coloniales, 1949.

Benedick, R.E., *Ozone Diplomacy: New Directions in Safeguarding the Planet*, Harvard University Press, Cambridge, Massachusetts, Enlarged Edition, 1998.

Brown, L., and E.P. Eckholm, *By Bread Alone*, Praeger Press, New York, 1974.

Callendar, G.S., The artificial production of carbon dioxide and its influence on temperatures, *Q. J. Roy. Meteor. Soc.*, 64, 223237, 1938.

Glantz, M.H. (ed.), *La Niña and Its Impacts: Facts and Speculation*. United Nations University Press, Tokyo, Japan, 2002.

Glantz, M.H. (ed.), *Societal Responses to Regional Climatic Change: Forecasting by Analogy*, Westview Special Study, Boulder, Colorado, 1988.

Glantz, M.H. (ed.), *Desertification: Environmental Degradation in and around Arid Lands*, Westview Press, Boulder, Colorado, 1977.

Glantz, M.H. (ed.), *The Politics of Natural Disaster: The Case of the Sahel Drought*, Praeger Press, New York, 1976.

Hare, K., Connections between climate and desertification, *Environ. Conserv., 4* (2), 82, 1977.

Huntington, E., *Civilization and Climate*, Yale University Press, New Haven, Connecticut, 1915, rev, ed. 1924, reprinted by University Press of the Pacific, 2001.

IPCC (Intergovernmental Panel on Climate Change), *Climate Change 2001: Impacts, Adaptation, and Vulnerability*, Contribution of Working Group II to Third Assessment Report, Cambridge University Press, Cambridge, UK, 2001.

IPCC, *Climate Change 1995: The Science of Climate Change*, Contribution of Working Group I to Second Assessment Report, Cambridge University Press, Cambridge, UK, 1996.

IPCC, *Climate Change: The IPCC Scientific Assessment*, Cambridge University Press, Cambridge, UK, 1990.

Kamarck, A.M., *The Tropics and Economic Development: A Provocative Inquiry into the Poverty of Nations*, The Johns Hopkins University Press, Baltimore, Maryland, 1976.

Markham, S.F., *Climate and the Energy of Nations*, Oxford University Press, London, 1944.

Myers, N., Environmental unknowns, *Science*, 269, 358360, 1995.

Ponte, L., *The Cooling*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1976.

Pye, L., *Aspects of Political Development*, Little, Brown and Co., Boston, Massachusetts, 1966.

Revelle, R.R., and H.E. Suess, Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric CO_2 during the past decades, *Tellus*, 9, 1827, 1957.

Sen, A., *Poverty and Famines: An Essay on Entitlement and Deprivation*, Oxford University Press, Oxford, UK, 1981.

Smith, J.B., N. Bhatti, G. Menzhulin, R. Benioff, M.I. Bodyko, M. Campos, B. Jallow, and F. Rijsberman (eds.), *Adapting to Climate Change: An International Perspective*, Springer-Verlag, New York, 1996.

Streets, D.G., and M.H. Glantz, Exploring the concept of climate surprise, *Global Environ. Chang.*, 10, 97107, 2000.

Weather Conspiracy: The Coming of the New Ice Age, Ballentine Books, New York, compiled by The Impact Team, 1977.

Wolde Mariam, M., *Rural Vulnerability to Famine in Ethiopia, 195777*, Vikas Publishing House, New Delhi, 1984.

CHAPTER 38

HOUSEHOLD FOOD SECURITY AND COPING WITH CLIMATIC VARIABILITY IN DEVELOPING COUNTRIES

THOMAS E. DOWNING AND YOLANDE STOWELL

1 INTRODUCTION

Climate affects food security in two distinct ways. Primarily, climate in association with soils, terrain, and vegetation is a resource that influences potential agricultural production. Agricultural productivity, in turn, provides income to individuals and households, leads to investment in infrastructure, and fuels the regional economy.

Climate also includes the hazards of drought, flood, windstorms, hail, and temperature extremes. Such climatic hazards lead to direct losses of income, infrastructure, and even lives. Indirect effects of hazards include aversion to investment because of high risk, a lack of infrastructure, and stagnation of the regional economy. Drought is the leading climatic hazard for vulnerable populations in most developing countries, although flood risk is critical in South Asia and China, and in recent years to some parts of Central America.

Conceptions of food security have evolved over the past few decades, and it has increasingly been recognized that "much more was involved in food security than just climate" (Glantz, 1997). Since the United Nations (UN) World Food Conference in 1974, there have been three major shifts in food security thinking: changing the scale from global to household and individual, broadening the focus from food alone to long-term resilience of livelihoods, and diversifying from objective indicators such as target levels of consumption to more subjective perceptions of security (Maxwell, 1994). (See Box 1 for a summary of food security definitions.)

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

Box 1 Definitions of Food Security

Food security is defined in its most basic form as access by all people at all times to the food needed for a healthy life. Achieving food security has three dimensions; first, it is necessary to ensure a safe and nutritionally adequate food supply both at the national level and at the household level. Second, it is necessary to have a reasonable degree of stability in the supply of food both from one year to the other and during the year. Third, and most critical, is the need to ensure that each household has physical, social, and economic access to enough food to meet its needs. This means that each household must have the knowledge and ability to produce or procure the food that it needs on a sustainable basis. In this context, properly balanced diets that supply all necessary nutrients and energy without leading to overconsumption or waste should be encouraged. It is also important to encourage the proper distribution of food within the household, among its members.

The right to an adequate standard of living, including food, is recognized in the *Universal Declaration of Human Rights*. Food security should be a fundamental objective of development policy as well as a measure of its success. Household food insecurity affects a wide cross section of the population in both rural and urban areas. The food-insecure socioeconomic groups may include: farmers, many of them women, with limited access to natural resources and inputs; landless laborers; rural artisans; temporary workers; homeless people; the elderly; refugees and displaced persons; immigrants; indigenous people; small-scale fishermen and forest dwellers; pastoralists; female-headed households; unemployed or underemployed people; isolated rural communities; and the urban poor. Increasing the productivity and incomes of these diverse groups requires adopting multiple policy instruments and striking a balance between short-term and long-term benefits. The choice of policies must be attuned to the characteristics of a country's food security problem, the nature of the food-insecure population, resource availability and infrastructural and institutional capabilities at all levels of government and communities. Breast-feeding is the most secure means of assuring the food security of infants and should be promoted and protected through appropriate policies and programs.

Source: International Conference on Nutrition, Plan of Action, Rome, Italy, 11 December 1992, from *www.brown.edu/Departments/World_Hunger_Program/hungerweb/intro/food_security.html*; compiled by: Nancy B. Leidenfrost, National Program Leader, Extension Service, USDA.

Four epochs can be distinguished in the history of food security of developing countries. Precolonial societies were dominated by rural, *self-provisioning* economies with various forms of organization generally based on ethnic affiliation. Trade linked remote areas with overseas territories, but production was primarily agricultural—cultivation, livestock rearing, hunting, and gathering. Resilience in the face of

drought depended on the ability to store food, either within the household or among kin.

Colonialism sundered traditional land tenure and governance. New forms of vulnerability were created in the transition from self-provisioning to a mixture of local and national governance. The *political economy of colonialism* determined access to land and to famine relief in the case of a drought.

With independence, the state continued to dominate economic systems. However, *weak national economies and political systems* were often unable to respond to famines or indeed to ameliorate widespread impoverishment. The catastrophic famines of the 1970s illustrate the enhanced vulnerability resulting from international and national political conflicts, terms of trade that failed to promote development, and hindrances in information.

Current vulnerability might be labeled *interdependence*. National states no longer dominate in famine early warning systems and food interventions. Market forces predominate in determining access to resources. There is some sense of progress in the ability of international aid organizations to monitor and prevent famines. However, many conditions of food insecurity persist in endemic poverty and in countries and regions isolated for political reasons (see Box 2).

Box 2. World Food Security

The World Food Summit of 1996 reviewed the state of world food security. Gains in agricultural productivity and economic growth over the last 30 years has led to an 18% growth in world per capita food supply. Average per capita dietary energy supply (DES) has grown from 2440 to 2720 calories/capita/day from 1969–1971 to 1990–1992. Aggregate figures, however, do not show the full picture, and hunger persists. In the period 1990 to 1992 an estimated 840 million people remained chronically undernourished, having access to less than 2700 calories per day. In addition, large numbers suffer from micronutrient deficiencies caused by dietary inadequacies, an example being the estimated 1.6 billion suffering from iodine deficiency. In absolute terms food aid has declined, while the increasing number and complexity of emergencies has resulted in a growing proportion (from 30 to 50% in two decades) of total food aid being targeted relief and development food aid.

Source: FAO, Technical Background Documents, World Food Summit, Rome 13–17 November, 1996; *http://www.fao.org/wfs/final/e/list-e.htm*.

In this section three case studies set the scene for further discussions of vulnerability and food security from a household perspective. These case studies indicate the range of situations regarding household food security and coping with climatic variability. In addition, recent research on climate prediction and its role in alleviating food security is summarized.

2 CASE STUDIES OF VULNERABILITY AND COPING

Changing Vulnerability in Central and Eastern Kenya

The failure of the first rains in 1984, following poor second rains in 1983, triggered a serious food crisis in central and eastern Kenya. The drought illustrates diverse impacts and responses in different environments.

The study area comprises six districts in central and eastern Kenya: Kiambu, Murang'a, Embu, Machakos, and Kitui, spanning five agroecological zones (Table 1) [see Downing (1988) and Downing et al. (1989*b*), for details]. The highland zones (I and II) are suitable for coffee and tea as cash crops, along with maize as the staple food crop. Farm sizes are small due to pressure from population growth. In the middle zone (III) maize is the dominant crop, although farms are also relatively small. In the drier zones (IV and V), farms are larger, with more livestock and grazing. Maize is still common, but millet and sorghum are more suitable. The arid ranching zone (VI) is dominated by extensive land uses.

Several long-term trends have affected food security in the region:

- Maize, although more drought prone, has replaced sorghum and millet.
- Farmers have less variety in their cropping systems, partly due to smaller holdings and the reduced availability of common lands.
- Markets for crop inputs, trade, and employment have penetrated the region and are well connected with Nairobi and the coast.

Food production in the six districts declined in 1984. Maize and bean harvests for the 1984 long rains were less than 20% of the harvest from the 1985 long rains. The effect of the drought on livestock was equally severe. Of 565 households surveyed in the study area, over a third had slaughtered, sold, or lost cattle, and a quarter had slaughtered, sold, or lost goats. The average number of cattle per household declined from over 4 in April 1984 to 2.4 in January 1985. The largest declines were in the lower agroclimatic zones. However, famine was averted.

The sources of household income changed (Table 2). Whereas 60% of the households usually have income from sales of farm produce, only 40% did so during the drought. With less food produced, there was less surplus after household consumption. Income from casual labor and businesses decreased as well. A general recession in the rural economy lowered casual employment opportunities and earnings. Income from remittances and permanent employment tended to increase or remain at their usual levels. Although livestock sales increased, the market collapsed and little income was realized.

The principal response by the government was to import yellow maize. Food prices in local markets increased during the drought, but less than would have been the case in the absence of abundant yellow maize imports. Prices of white maize and beans doubled between early 1984 and late 1984, while the imported yellow maize prices remained about the same. In most markets, food was available; complete dearth of food for extended periods was rare.

Households generally survived the food crisis by purchasing their food: participation in the monetary economy reduced their vulnerability to drought. Households are largely self-sufficient in average years with current farming practices and land holdings. A slight surplus can be produced in the humid zones (II), where maize grows well. The larger holdings in the semiarid zones (V) also produce a surplus in years of average to good weather.

Drought affects each zone, with increasing intensity from zones II to V. In a severe drought, production is half of household requirements in zone II, while there is little or no production in zones IV and V. Neither improvements in yields nor adoption of present drought-resistant crops will significantly improve self-sufficiency in drought years. In good and average years, production can be dramatically increased with fertilizers and other inputs. Households in the humid zones could be largely self-sufficient in drought years. But in the drier zones, little improvement in production can be expected in moderate or severe droughts.

In contrast, storage of surplus food is a potentially effective drought-coping strategy, depending on good production during above-average years. However, food storage has a cost. In central Kenya, average food storage in the 1980s was equivalent to 40 to 100 days of consumption, and twice that in eastern Kenya, which is drier and more prone to drought. These low levels of food storage indicate a preference to sell surplus food in order to pay for school fees, medical expenses, and for food not grown on the farm.

Rangeland Management in Botswana

Drought occurs, on average, once every 7 years in Botswana. As elsewhere in Africa, drought has been a common cause of food insecurity. Two studies highlight the confluence of drought and the political economy.

Was the 1979 to 1987 drought more severe than preceding dry periods? Years after the 1979 to 1987 drought in Botswana, the government had not withdrawn relief to many areas and had in some areas expanded relief efforts. Solway (1994) argues that non-meteorological factors are critical to the failure of much of rural Botswana to return to "normal" after 1987. The differential effects of drought—the distribution of both benefits and disadvantages among various classes, races, and among men, women, and children—cannot be sought solely in meteorological factors.

The drought made legitimate a shift of dependency from the extended family to the state and subsequently a greater dependency on the state. Traditional patterns of food security, which permitted semi-independent production on the part of the poorer majority, were significantly eroded during the drought. They have not been revived. In the case of Kalahari villagers, access to draft animals through a chain of entitlements based on kinship relations was replaced by state social security introduced during the drought and then maintained. This shift in dependency was favored by the rural elite, who saw an opportunity to consolidate wealth with the commodification of agriculture and privatization of production occurring in rural Botswana. This shift was also supported by the government, which viewed traditional patron–

TABLE 1 Agroclimatic Zones in Kenya[a]

Zone	R/E_0 (%)	Growing Season (days)	Dry Matter (mt/ha)	Population Density (km^2)	Major Crops	
					Food	Commercial
I	>80	365	>30	333	Beans, maize, potato	Coffee, dairy, tea
II	65–80	290–365	20–30	468	Beans, maize, potato	Dairy, coffee
III	50–65	235–290	12–20	275	Beans, cassava, cow peas, green grams, maize, pigeon peas, sorghum, millet	Cotton, fruit, tobacco
IV	40–50	180–235	7–12	111	Beans, cassava, cow peas, green grams, maize, pigeon peas, sorghum, millet	Cotton, fruit, sunflower, tobacco
V	25–40	110–180	3–7	36	Millet, sorghum	Cotton, livestock, sisal, sunflower
VI	<25	<110	<3	<5		Livestock, sisal

[a] R/E_0 is the ratio of rainfall to potential evaporation. Population density is for 1979.

Source: Sombroek et al. (1982), Downing, Lezberg et al. (1989b) and Jaetzold and Schmidt (1983).

client relations as "backward" and favored "an ideal of individualized nuclear family production units which functioned independently of one another but in conjunction with the state" (Solway, 1994, p. 491). It brought the rural sector in line with modern systems of taxation, land registration, and government regulation.

With the introduction of "welfare" and the loss of access to the local means of production, poorer rural residents were no longer able to farm, and overall rural agricultural production dropped, despite the fact that bumper harvests were reported in several villages during the study period. Falling production statistics provided further "evidence" for the severity of the drought and justified the further expansion of government relief programs.

In Solway's conceptualization, the drought was a revelatory crisis, arising from "structural contradictions" between traditional and modern, bringing latent societal tensions to the surface, and providing a context for the accelerated change of the bases of social reproduction. These structural contradictions were simultaneously

TABLE 2 Sources of Household Income during the 1984 Drought in Kenya

	Agrochemical Zone					
Income Source	I	II	III	IV	V	Total
Farm produce	90[a]	71	50	67	50	62
	98	113	68	28	28	65
Livestock	14	17	14	19	57	29
	100	118	164	200	107	88
Agricultural casual labor	27	25	31	35	31	31
	85	96	94	97	68	87
Nonagricultural casual labor	20	10	18	25	37	23
	100	100	89	84	81	87
Businesses	10	14	15	12	22	15
	70	57	100	125	91	93
Remittances	35	25	41	46	41	39
	86	100	90	109	95	97
Permanent employment	24	24	32	29	35	30
	92	79	106	107	83	97

[a]The first number in each cell is the percent of households for whom the income source is a usual source of income. The second number is the ratio (in percent) of households who had the income source during the 1984 drought (April to December 1984) compared to the number of households who had it as a usual source. Thus over all of the zones, 62% of the households usually have farm produce as a source of (cash) income. During the drought months, only 65% of these households had income from their farm produce.

Source: Anyango et al. (1989).

revealed and concealed during the drought; the discourse of crisis allowed them to be hidden from view by claims that "exceptional" circumstances prevailed. Discourse here served a dual purpose: Not only did it obscure deeper processes at work, it also impelled and made legitimate innovation with normative codes and regulatory, management, and institutional frameworks.

Other research in the Kalahari area of Botswana, investigating the effects of the 1975 tribal grazing lands policy (TGLP), also notes the importance of the social and political dimensions to food security (Thomas and Sporton, 1997b). Studies focused on three areas: study area 1 in eastern Kalahari with tree and bush savannah, study area 2 in Ncojane with dryer bush savannah, and study area 3 in Tshabong, the driest area of arid shrub savannah. It was found that the areas with the most flexible approaches to the TGLP were best able to overcome impacts of environmental variability, notably the drought of 1994 to 1995. The adaptive strategies listed in Table 3 were only possible if some elements of the TGLP, such as fencing of ranches, were not enforced. Study area 2, therefore, enjoyed better livelihood security than study area 1, where TGLP was applied most rigorously, despite the fact that on ecological grounds area 1 offered better opportunities.

The style of management and relationship between ranch lessees and workers also has a major effect on food security. Study area 2 had a more participatory, flexible management style, which minimized hardship (see Box 3).

TABLE 3 Adaptive Strategies Employed in Study Area 2

By livestock holders/ranch lessees
Fences dropped between ranches/paddocks to increase grazing range
Temporary removal of livestock to distant cattleposts
Pooling of resources with neighboring lessees
By ranch residents/other rural groups
Settlement on (temporarily) abandoned ranches
Gathering (and hunting) across neighboring ranches
Move residence to ranch nearest to service center, where drought relief may be available

Source: Thomas and Sporton 1997b.

Box 3. Perceptions of Food Security

Interview from study area 1 (male in his fifties):

"I don't get any food rations from the owner. I just wake up hungry every morning. I help myself live by milking the cows and drinking the milk. I don't know why the owner of the ranch is treating me this way. I don't know if he hates me because of my ethnic group, because I'm a Moswara . . . When I was a young man I used to be satisfied every day . . . But now that I am working for someone I don't eat at all."

Interview from other study area 2 (male in his fifties):

"I came here to look after the ranch-owner's cattle. I'm not paid for doing the work. He doesn't pay me, he just buys me food. Not being paid is not a big problem. I'm OK as things are. When I need something I just tell the ranch owner and he buys it for me. . .The owner is a good man I'm happy to work for him."

Source: Thomas and Sporton (1997a).

Displaced Populations in Sudan

Sudan has a large population of displaced people who have left their original villages and moved elsewhere, often to urban settlements. The Commission of Displaced (COD) estimates that 3.5 million people had been displaced by 1991, of which 0.6 to 1.5 million were in the greater Khartoum area (Kuch, 1993). The major factors forcing the migrations were civil war in the south, which erupted during the dry summer of 1983, and the 1984 drought that resulted in the worst famine to hit northern Sudan in the past 100 years (Kuch, 1993). The underlying causes for displacement go well beyond escaping the impact of drought. A mixture of socio-economic and environmental pressures is responsible, partly resulting from the

TABLE 4 Crisis-Induced Migration Model (CIM)

1.	Normality (pre-migration situation)
2.	Emergency situation
3.	Local reaction chain
4.	Migration
5.	Sanctuary phase (seeking refuge of a temporary nature)
6.	Settlement phase
7.	Return phase

Source: Elnur et al. (1993, p. 50); developed by the UN Emergency Unit.

dominant development model that for years ignored the traditional sector (Elnur et al., 1993).

Most studies of food security and survival strategies focus on the premigration situation and regard the act of migration itself as the final coping strategy (see Dagnew 1995, p.109). To investigate the food security strategies of displaced populations, it is necessary to look at stages beyond migration (see Table 4). It is important to note that where war is a predominant forcing factor the expected premigration sequence of events may not be followed. The speed and forced nature of this type of migration increases the vulnerability of the displaced as they are likely to have lost all assets, leaving nothing with which to start their postmigration life (Elnur et al., 1993).

The displaced populations are not a homogeneous group, which is reflected in the diversity of the survival strategies adopted (see Table 5). These strategies are a function of many socioeconomic variables such as the predisplacement economic activity of the people. For example, many of those displaced from the south were cattle herders or subsistence farmers with few professional skills. The sex, number, and age of household members are also major determinants of livelihood strategies; 35% of displaced families are female headed (Kuch, 1993). Even within groups with the same basic means of livelihood, tremendous differences can occur; poorer families may have trouble obtaining credit to enable them to diversify their activities.

A study by Kuch (1993) also investigated the food situation of the displaced in Khartoum and found them to be highly vulnerable. People without access to land cannot produce their own food, making them heavily dependent on the market economy. Monetary income is, therefore, crucial to food security in these situations, and people even resort to illegal methods, such as the sale of alcohol, which is punishable with lashes and fines and up to a year imprisonment (Kuch, 1993). Inability to purchase even a single meal a day was frequently experienced. Malnutrition among those under 5 years old was found to double within a year in most settlements. Rations introduced by the Sudan Council of Churches' (SCC) Primary Health Care Program (PHCP) to supplement their diets had to be extended to entire families because sharing of the rations meant that the health status of the under-5-year-olds often did not improve (Kuch, 1993).

TABLE 5 Responses Adopted by Displaced Southern Sudanese to Secure Their Livelihood

Seeking shelter with relatives already residing in planned residential areas
Working as guardians at construction sites
Departure of young men for mechanized agricultural schemes
Retail selling of different consumer items
Selling of rationed supply items
Child labor at markets
Domestic work of women and children, mainly washing and cleaning other people's houses
Working as daily workers at construction sites and industries
Production and distribution of local alcoholic beverages
Depending on government and nongovernmental organisations (NGOs) aid
Begging and collecting of rubbish
Robbery and other illicit activities, including prostitution

Source: Elnur et al. (1993, p. 51).

Government support for food-insecure migrants comes in the form of the "essential commodity distribution card" system, which subsidizes products such as sorghum, sugar, tea leaves, vegetable oil, batteries, matches, and washing soap. For the first 9 months of the scheme, unregistered quarters, including displaced settlements, were not given cards. Even when the displaced were given cards, their purchasing power limited them because, despite subsidies, the commodities' prices remained more than many could afford (Kuch, 1993). Relief food was distributed by agencies under government surveillance.

Many migrants will not achieve truly sustainable livelihoods while in camps and settlements. They await the return to their home area as the only solution to their survival problems (Yath, 1993). Further work is needed in order to assist policy decisions such as planning what type of food distribution (e.g., commercial or relief systems) is best in settlements. Also, more attention needs to be directed at assisting people to develop and expand their own coping strategies.

3 APPROACHES TO COPING, CAPACITY, AND VULNERABILITY

The three case studies highlight diverse situations and ways in which households have coped with climatic variations and other stresses. A rich literature now exists on coping strategies, capacity to withstand shocks and stress, and more generally vulnerability. A few observations on vulnerability frameworks are now presented as an introduction to the following section, which specifically addresses the intersection of vulnerability and climate prediction.

Vulnerability can be defined as the "degree of loss resulting from a potentially damaging phenomenon" (UNDHA, 1992, p. 63) or "the insecurity of the well-being of individuals, households, or communities in the face of a changing environment" (Moser, 1996, p. 2).

Critical questions are: What determines the relationship between a hazard and its effects? Who is vulnerable and why? These questions require a broader analysis of vulnerability. This amplification of vulnerability stems from the literature on development and livelihood security [e.g., see Chambers (1989) and Dow and Downing (1995)] rather than the more circumscribed work on disaster relief. As such, it begins to place vulnerability in the wider structures of political ecology.

Particular vulnerabilities are the conjuncture of social, economic, and political structures. Anderson and Woodrow (1989) chart vulnerability and capability related to physical/material resources, social/organizational relations, and motivational/attitudinal aspects. Bohle et al. (1994) suggest a tripartite causal structure of vulnerability (Fig. 1) based on the human ecology of production, expanded entitlements in market exchanges, and the political economy of accumulation and class processes. Vulnerability per se is best viewed as "an aggregate measure of human welfare that

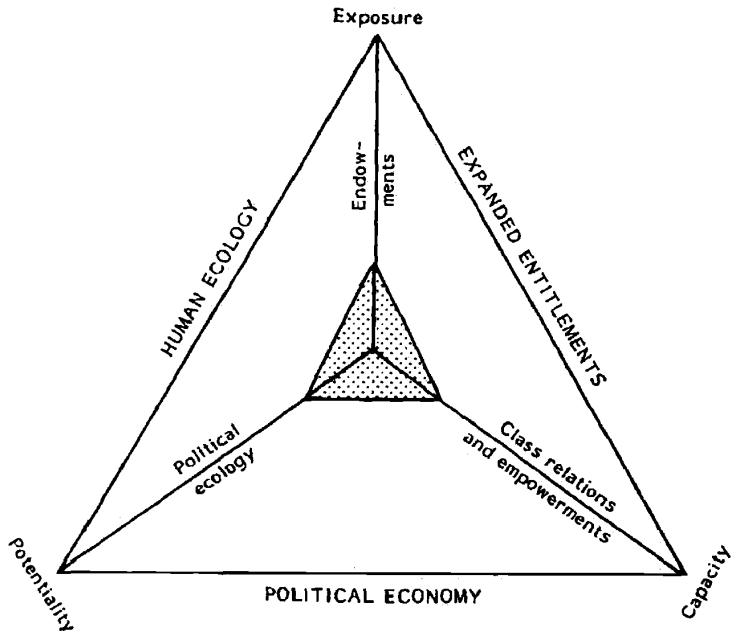


Figure 1 Three dimensions of vulnerability. The fundamental processes that determine vulnerability are implied in the conjuncture of the human ecology of production, exchange entitlement, and political economy. Vulnerable groups can be located in different sectors of the triangle. For instance, subsistence farmers would be more dependent on their land and labor resources than on market exchanges. The destitute and refugees are closely tied to the political economy of aid. The urban poor are dependent on what they can earn in informal markets (after Bohle et al., 1994).

integrates environmental, social, economic and political exposure to a range of harmful perturbations" (Bohle et al., 1994, pp. 37–38). In one of the fullest treatments of vulnerability and disasters, Blaikie et al. (1994) regard vulnerability as a product of such characteristics as ethnicity, religion, caste membership, gender, and age that influence access to power and resources (Fig. 2). One application of the concept of vulnerability is in the U.S. Famine Early Warning System, which monitors food crises in Africa (Fig. 3).

Regardless of the nuance of vulnerability frameworks, key concepts are:

- Vulnerability is a relative measure. The analyst, whether the vulnerable themselves, external aid workers, or various societies that include both the vulnerable and interventionists, must define what is a critical level of vulnerability.
- Everyone is vulnerable, although their vulnerability differs in its causal structure, its evolution, and the severity of the likely consequences.
- Vulnerability relates to the consequences of a perturbation, rather than its agent. Thus people are vulnerable to loss of life, livelihood, assets, and income rather than to specific agents of disaster, such as floods, windstorms or technological hazards. This focuses vulnerability on the social systems rather than the nature of the hazard itself.
- The locus of vulnerability is the individual related to social structures of household, community, society, and world system. Places can only be ascribed a vulnerability ranking in the context of the people who occupy them.

These concepts of vulnerability shift the focus of vulnerability away from a single hazard to the characteristics of the social system. Vulnerability is explicitly a *social* phenomenon, a threat to a human value system. Places and ecosystems can only be termed vulnerable if we ascribe human value to them.

Vulnerability changes over time, incorporating social responses as well as recurrences of hazardous events. Bohle et al. (1994) captures the dynamic nature of vulnerability (Fig. 4). In this illustration, vulnerability begins to increase at the end of the first year, reaching a crisis at 30 months. Here the outcome of the crisis is uncertain. In a resilient society with appropriate interventions, recovery and mitigation can bring vulnerability back down to baseline (or lower) levels. Unmitigated, or in conjunction with another event such as civil strife following drought, the crisis may become a disaster. Or, some groups and communities may continue in crisis, on the edge of disaster. Social groups vary in the structure of their vulnerability. For example, the rural landless (without nonagricultural incomes) are typically more sensitive to food shortages, with less on-farm storage and buffering capacity than smallholders. Thus, the trajectories shown here may be sharper and the outcome different for different groups, even in the same region.

Specific groups of vulnerable peoples can be defined. While the precise boundaries of vulnerability vary between cultures and environments, the common catalog often starts with the characteristics of individuals:

PROGRESSION OF VULNERABILITY								
ROOT CAUSE	⇒	*DYNAMIC PRESSURES*	⇒	**UNSAFE CONDITIONS**	⇒	*DISASTERS*	⇐	HAZARDS
Limited access to • Resources • Structures • Power **Ideologies** • Political systems • Economic		**Lack of** • Institutions • Training • Skills • Investment • Markets • Press freedom • Civil society **Micro-forces** • Population • Urbanization • Arms expenditure • Debt repayment • Deforestation • Soil degradation		**Fragile physical emnvironment** • Dangerous locations • Unprotected structures **Fragile local economy** • Livelihoods at risk • Low income **Vulnerable society** • Groups at risk • Little capacity to cope **Public actions** • Lack of preparedness • Endemic disease		**RISK** (equals) **HAZARD** (plus) **VULNERABILITY**		**Earthquake** **Wind storm** **Flooding** **Volcano** **Landslide** **Drought** **Virus and pest** **Heatwave**

Figure 2 Structure of vulnerability and disasters. Dynamic pressures are processes that translate social, political, and economic structures in relation to specific types of hazards into particular forms of insecurity. Regional or global pressures such as rapid population growth, urbanization, war, foreign debts, epidemic disease, export promotion, etc. have effects on the local or regional level. Some of these pressures have a universal character, others are specific to a certain region or society. Unsafe conditions reflect situations and circumstances specific to a region and time, in conjunction with a particular group of people. They reflect specific forms of vulnerability related to specific hazards (Blaikie et al., 1994).

Level of Vulnerability	Conditions of vulnerability	Typical Coping Strategies and/or Behaviours	Interventions to Consider
SLIGHTLY VULNERABLE	**Maintaining or Accumulating Assets**	**Assets/resources/wealth:** either accumulating additional assets/resources/wealth or only minimal net change (normal "belt-tightening" or seasonal variations) in assets, resorces or welth over a season/year, i.e., coping to minimize risk	**Development Programs**
	Maintaining Preferred Production Strategy	**Production Strategy:** any changes in production strategy are largely volitional for perceived gain, and not stress related	
MODERATELY VULNERABLE	**Drawing-down Assets**	**Assets/resources/wealth:** coping measures include drawing down or liquidating less important assets, husbanding resources, minimizing rate of expenditure of wealth, unseasonable "belt-tightening" (e.g., drawing down food stores, reducing amount of food consumed, sale of goats or sheep)	**Mitigation and/or Development: Asset Support** (release food price-stabilization stocks, sell animal fodder at "social prices", community grain bank, etc.)
	Maintaining Preferred Production Strategy	**Production strategy:** only minor stress-related change in overall production/income strategy (e.g., minor changes in cropping/planting practices, modest gathering of wild food, inter-household transfers and loans, etc.)	
HIGHLY VULNERABLE	**Depleting Assets**	**Assets/resources/wealth:** liquidating the more important investment, but not yet "production" assets (e.g., salw of cattle, sale of bicycle, sale of professionals such as jewelry)	**Mitigation and/or Relief: Income and Asset Support** (Food-for-work, Cash-for-work, etc.)
	Disrupting Preferred Production Strategy	**Production Strategy:** coping measures being used have a significantly costly or disruptive character to the usual/preferred household and individual life-styles, to the environment, etc. (e.g., time-consuming wage labor, selling firewood, farming marginal land, labor migration of young adults, borrowing from merchants at high interest rates)	
EXTREMELY VULNERABLE	**Liquidating Means of Production**	**Assets/resources/wealth:** liquidating "production" resources (e.g., sale of planting seed, hoes, oxen, land, prime breeding animals, whole herds)	**Relief and/or Mitigation: Nutrition, Income and Asset Support** (food relief, seed packs, etc.)
or			
AT-RISK	**Abandoning Preferred Production Strategy**	**Production strategy:** Seeding non-traditional sources of income, employent, or production that preclude continuing with preferred/usual ones (e.g., migration of whole families)	
FAMINE	**Destitute**	**Coping Strategies Exhausted:** no significant assets, resources, or wealth; no income/production	**Emergency Relief** (food, shelter, medicine)

Figure 3 Vulnerability Matrix for the U.S. Famine Early Warning System. Vulnerability is portrayed as a progression from slightly vulnerable to famine. At each level of vulnerability, households pursue different strategies—from production to survival. Consequently, different forms of intervention are warranted for different levels of vulnerability, from targeted development assistance to supporting coping strategies and ultimately emergency food relief [U.S. Famine Early Warning System (FEWS), 1992].

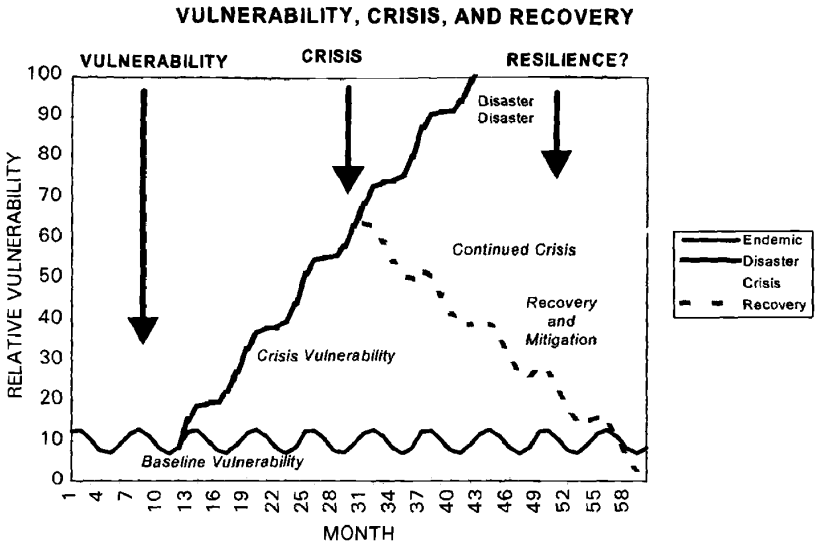


Figure 4 Dynamic vulnerability during a food crisis (after Bohle et al., 1994).

- Women, especially those with special nutritional needs during and after pregnancy
- Children, who are less resilient in terms of nutrition or who may already be malnourished
- Elderly, who may suffer from a lack of mobility and less mental awareness
- Disabled and disease stricken, who have special needs and require routine assistance for survival

At the household level, vulnerability may be delineated by socioeconomic class and means of securing a livelihood: In rural areas:

- Smallholder agriculturalists may be resource poor with limited access to land and labor, in marginal lands, with varying degrees of empowerment and access to emergency and development assistance.
- Pastoralists often have little empowerment to development resources, yet operate in regions with pronounced climatic hazards. However, they often attract international assistance during a disaster.
- Landless laborers relying on casual employment are often at the margin of poverty with little ability to accumulate savings or to invest in more productive activities.
- Destitute peoples have been forced out of productive activities, often because of ill health and old age in addition to being impoverished through natural

disasters and other causes. Where the destitute migrate to urban centers, they may have more opportunities for assistance and work, although this depends on the nature of the receiving society.

In urban areas:

- Unemployed destitute in urban areas may be incorporated into social welfare systems (often informal), but suffer significantly in times of disaster if the numbers become too large and if relief fails to target their pressing needs.
- Underemployed poor people, comparable to landless laborers, are on the margins of survival. A slow deterioration in the economy can affect this group, often leading to a major but largely hidden crisis.
- Refugees are the most visible vulnerable population, usually swelling in numbers after a disaster. They may also be vulnerable to further hazards, for instance, while attempting to return to their homes and occupations or in camps with inadequate protection against floods, heat, and frost, among other hazards. Yet, this group tends to benefit from its visibility and various formal channels of assistance.

Finally, community characteristics can be enumerated that place vulnerable social groups at risk from specific types of hazards. Such vulnerability includes, for example:

- Building location, design, and standards that determine the resilience of homes, workplaces, and community meeting places
- Occupancy patterns and who is in substandard buildings, when, for how long, etc.
- Transport and mobility, for example, the pathways between home and work may cross hazardous regions and access to safe areas such as cyclone shelters in Bangladesh
- Health, water, power, and communication infrastructure that sustain life as well as provide channels for relief assistance.

4 COPING AND CLIMATE PREDICTION

How does the broad range of vulnerability and capacity in Africa relate to emerging skills in climate prediction? Recent developments in seasonal forecasting, especially for the tropics (e.g., Chen et al., 1995), have drawn attention to the opportunity for appropriating such forecast information into drought management systems and other natural resource operations (e.g., Gibberd et al., 1996) (Table 6). Africa is one of the potential beneficiaries of such improved forecasts.

The wide range of effects that climate, and particularly drought, can have is discussed in Glantz's work on El Niño (Glantz, 1996, pp. 145–148). It is emphasized

TABLE 6 Qualitative Assessment of Current Status of Long-Lead Climate Forecasts

	Operational Status Depending on Region			
	Untried	Experimental	Pre-operational	Operational
Multiyear (climate)	Most areas	Global/ hemispheric-scale	—	—
Seasonal	Some equatorial and high-latitude areas	Many areas	Certain promising areas including southern Africa	Parts of the United States, Australia, and a few other suitable areas
Within season	—	Many areas including southern Africa	Some developed economies	Certain well-researched areas (United States, Europe)

Source: Gibberd et al., (1996).

that "weather" does not only affect crop yield but also land quality, on-farm storage, labor migration, rates of urbanization and rural population growth, use of inputs such as fertilizer, farm income, farmers' skill and experience, and so forth. The utility of reliable long-range forecasts, therefore, could be enormous, not just for earlier warning of need for emergency aid but also for ongoing food security (Table 7). Policymakers and farmers alike should benefit.

Seasonal forecasts are already being used in some parts of Africa, for example, in predicting maize yields in Zimbabwe (Cane et al., 1994). For agriculture and water resource management the benefits could be quite extensive, altering the entire basis of economic planning in Africa. Most farmers would benefit from seasonal forecasts, although lead time and reliability will be important issues (Table 8). Several studies have analyzed costs of El Niño events, for example, and predict that considerable savings could be made if accurate warnings of the onset of the phenomenon could be used. The 1991–1992 El Niño-related drought in southern Africa was estimated to cost the U.S. government $800 million in responses to the phenomena (Farmer 1997).

5 CONCLUSIONS

The three case studies focus on drought and associated famines in various parts of Africa over the past decade or so. Droughts in 1983–1984 and 1991–1992 were both described as unusual or the worst to affect the subcontinent in the twentieth century (Rook, 1997). It is remarkable that, despite serious reductions in harvests, widespread hunger was averted, at least in 1991–1992, a fact largely attributed to the

TABLE 7 Users and Potential Applications of Climate Forecasts

	Potential Application of Forecast			
Type of User	Multiyear Forecasts	Seasonal Forecasts	Within Season	Benefits
Commercial producers	Capital and land investment	Acreage planted; planting dates; crop/variety selection; water management	Water management; application of inputs; harvest dates	Increased certainty and reduced risk; improved financial viability; long-term survival; enhancement of comparative advantage
Subsistence producers	Limited, possible diversification and off-farm savings	Planting dates; crop/variety selection	Limited	Improved food security in poor years; improved marketable surpluses in good years
Agricultural support services	Plant and capital investment; research and development priorities; location decisions; production strategies	Product selection; sales forecasts; pricing policy	Adjustments to marketing strategy	Improved financial viability; ability to respond better to farmers' requirements; recovery from drought
Agricultural extension services	Promotion of drought mitigation strategies; development of improved extension advice	Preparation of climate-specific extension advice to subsistence and smallholder producers	Specific adjustments to earlier extension messages and advice	Better extension service to subsistence and smallholder producers

Source: After Gibberd et al. (1996).

TABLE 8 Utility and Requirements for Seasonal Forecasts for African Agriculturists

Weather	Farm Type			Decision	Required Lead Time (months)	Required Precision/ Accuracy[a] (%)
	Subsistence	Transitional	Commercial.			
Drought	✓	✓	✓	Plant or not plant; choice of crops and tillage; contingency plans for livestock and water	3	90
Overall quality of the rainy season	✓	✓	✓	Choice of crops, crop varieties and tillage; irrigation planning to use impounded water efficiently; arrange seasonal credit	3	80
Onset of planting rains	✓	✓	✓	Timing of field operations; expectation of yield where correlated to planting date	0.5–1	80
Nature of early rains	✓	✓	✓	Whether to risk dry planting, depending on frontal, widespread rainfall or convective, isolated, discontinuous rainfall	0.5	80
Beginning of midseason drought		✓	✓	Choice of variety and planting date	3	80
Length of midseason drought		✓	✓	Choice of crop	3	60
Severity of midseason drought		✓	✓	Preparing to divert grain crops to fodder	3	60
End of rainy season	✓	✓	✓	Timing of harvest operations; possibility of late catchcrops; planning postharvest tillage	2	80

(*continued*)

TABLE 8 *(continued)*

Weather	Farm Type			Decision	Required Lead Time (months)	Required Precision/ Accuracy[a] (%)
	Subsistence	Transitional	Commercial.			
Amount of winter rains		✓	✓	Plan summer crops for optimum winter cereal crop; possibility of other winter crops	6	80
Distribution of winter rains		✓	✓	Level of inputs to invest in winter crop	1	60
First frost date	✓	✓	✓	Planting date for late planted crops; cut-off planting date for frost-sensitive crops	6	80
Last frost date			✓	Date of winter cereal planting to avoid frost at anthesis; planning spring plantings under irrigation	6	80
Frost frequency over winter			✓	Preparedness for frost on winter horticultural crops	1	40
Dry season severity	✓	✓	✓	Off-season farm capital development; livestock management	1	40
Dry season length	✓	✓	✓	Disposal of crop residues; fodder rationing to livestock; livestock mobility and sales	1	40
Above-normal summer temperatures			✓	Precautions in dairying and horticulture	3	6
Below-normal winter temperatures			✓	Precautions in small stock and horticulture	3	40

[a]Precision/accuracy: 100% completely reliable, 0% same as no forecast.

Source: Based on Gibberd et al. (1996).

rapid responses instigated by regional early warning systems (e.g., Betsill et al., 1997).

The effects of climate, and in particular climatic hazards, depend very much on the socioeconomic vulnerability of the population. The use of climate prediction is also related to vulnerability, both in terms of direct effect and in terms of how easy it is for a given group to access climate predictions and respond to them. Building institutional capacity to provide medium-term climate forecasts to enhance adaptive resource management in Africa would be a major step forward both in achieving present development aims and in preparing for climate fluctuations and change. Research on how to disseminate information and ensure it is applicable at the household level is also crucial.

Recently, increased attention has been paid to the El Niño Southern Oscillation (ENSO) phenomenon and its links with climatic hazards throughout the world, such as droughts in Africa (Glantz, 1997; Wolde-Georgis, 1997). A great deal is being done on the physical side of climate prediction; however, there is a "major gap in the application of research findings" (Glantz, 1997). Improved communications between the climate community and communities dealing with food security, such as famine early warning systems, are needed. For example, forecasters need to be explicit about the spatial and temporal resolution and confidence limits of predictions (Farmer, 1997).

However, an effective climate forecast and use system is not in itself sufficient to bring about increased food security. The forecasting process tends to focus on natural determinants of famine and tends to distract attention from other factors that shape societal and household vulnerability. "There is frequently a danger that forecasting can become an end in itself, detached from many of the social processes that give rise to hunger and starvation" (Tapscott, 1997).

Would better forecasts have altered the outcomes in Kenya, Botswana, and Sudan? The answer depends on how the political economy would have adapted to widespread dissemination of forecasts (and other data on climate and production). Good forecasts, in each case, would not by themselves have been sufficient to ensure early responses, to bolster sustainable livelihoods, and to prevent vulnerable populations from being displaced.

There is a need to ensure that agricultural development occurs in such a way that longer-term sustainability, and not short-term production maximization, is the aim (Rook, 1997). Climate predictions can help in this aim, but other social, economic and political factors must also be considered.

REFERENCES

Anderson, M. B., and P. J. Woodrow, *Rising from the Ashes: Development Strategies in Times of Disaster*, Westview, Boulder, CO, 1989.

Anyango, G. J., T. E. Downing, et al., Drought vulnerability in central and eastern Kenya, in T. E. Downing, K. W. Gitu, and C. M. Kamau (Eds.), *Coping with Drought in Kenya: Local and National Strategies*, Lynne Rienner, Boulder, CO, 1989, pp. 169–210.

Bakhit, A. H., Low-salary government employees in Khartoum: Strategies and mechanisms of survival of the urban poor, in H. G. Bohle, T. E. Downing, J. O. Field, and F. M. Ibrahim (Eds.), *Coping with Vulnerability and Criticality: Case Studies on Food-Insecure People and Places*, Saarbrucken, Fort Lauderdale, FL, Breitenbach, 1993, pp. 81–96.

Blakie, O., T. Cannon, et al., *At Risk: Natural Hazards, People's Vulnerability, and Disasters*, Routledge, London, 1994.

Bohle, H. G., T. E. Downing, et al., Climate change and social vulnerability: Towards a sociology and geography of food insecurity, *Global Environ. Change*, *4*(1), 37–48, 1994.

Cane, M. A., G. Eshel, et al., Forecasting Zimbabwean maize yield using eastern equatorial Pacific sea surface temperature, *Nature*, *370*, 204–205, 1994.

Chambers, R., Vulnerability, coping and policy, *IDS Bull.*, *20*(2), 1–7, 1989.

Chen, D., S. E. Zebiak, et al., An improved procedure for El Niño forecasting: Implications for predictability, *Science*, *269*, 1699–1702, 1995.

Dagnew, E., Differential socio-economic impact of food shortages and household coping strategies: A case study of Wolaita District in southern Ethiopia, *Afr. Devel.*, *20*(1), 89–124, 1995.

Dow, K., and T. E. Downing, Vulnerability research: Where things stand, *Human Dimensions Q. 1*(3), 3–5, 1995.

Downing, T. E., *Climatic Variability, Food Security and Smallholder Agriculturalists in Six Districts of Central and Eastern Kenya*, Department of Geography, Clark University, Worcester, MA, 1988, p. 262.

Downing, T. E., K. W. Gitu, and C. M. Kamau (Eds.), *Coping with Drought in Kenya*. Lynne Rienner, Boulder, CO, 1989a.

Downing, T. E., S. Lezberg, et al., Population change and environment in central and eastern Kenya from 1969 to 1979, *Environ. Conservation*, 1989b.

Elnur, I., F. Elrasheed, et al., Some aspects of survival: strategies among the Southern Sudan displaced people in Greater Khartoum, in H. G. Bohle, T. E. Downing, J. O. Field, and F. M. Ibrahm (Eds.), *Coping with Vulnerability and Criticality: Case Studies on Food-Insecure People and Places*, Saarbrucken, Fort Lauderdale, Breitenbach, 1993, pp. 45–58.

Farmer, G., What does the famine early warning community need from the ENSO research community? *Internet J. Afr. Stud.* (2), available on-line, 1997.

Gibberd, V., J. Rook, et al., *Drought Risk Management in Southern Africa: The Potential of Long Lead Climate Forecasts for Improved Drought Management*, Chatham Maritime, Natural Resources Institute, 1996.

Glantz, M. H., *Currents of Change: El Niño's Impact on Climate and Society*, Cambridge University Press, Cambridge, 1996.

Glantz, M. H. Summary: ENSO/FEWS discussions, *Internet J. of Afr. Stud.* (2), available on-line, 1997.

Glantz, M.H., M. Betsill, and K. Crandall, *Food Security in Southern Africa: Assessing the Use and Value of ENSO Information*, NOAA Project Report, ESIG/NCAR, Boulder, CO, 1997.

International, C. o. N., *Plan of Action. N. P. L.*, compiled by N. B. Leidenfrost, Extension Service, USDA, Rome, Italy, available on-line, www.brown.edu/Departments/World_Hunger_Program/hungerweb/intro/food_security.html, 1992.

Jaetzold, R., and H. Schmidt, *Farm Management Handbook of Kenya*, Ministry of Agriculture, Nairobi, 1983.

Kuch, P. J., Food situation among the displaced Sudanese in Khartoum State, in H. G. Bohle, T. E. Downing, J. O. Field, and F. M. Ibrahim (Eds.), *Coping with Vulnerability and Criticality: Case Studies on Food-Insecure People and Places*, Saarbrucken, Fort Lauderdale, Breitenbach, 1993, 33–44.

Maxwell, S., *Food Security: A Post-Modern Perspective*, Institute of Development Studies, 1994.

Moser, C. O. N., *Confronting Crisis: A Summary of Household Responses to Poverty and Vulnerability in Four Poor Urban Communities*, World Bank, Washington, D.C., 1996.

Rook, J. M. The SADC regional early warning system: Experience gained and lessons learnt from the 1991–92 Southern Africa drought, *Internet J. Afr. Stud.* (2), available on-line, http://www.dir.ucar.edu/esig/enso, 1997.

Solway, J. S., Drought as a "Revelatory crisis": An exploration of shifting entitlements and hierarchies in the Kalahari, Botswana, *Devel. Change*, *25*: 471–495, 1994.

Sombroek, W. G., H. M. H. Braun, et al., *Exploratory Soil Map and Agroclimatic Zone Map of Kenya*, Kenya Soil Survey, Nairobi, 1982.

Tapscott, C., Is a better forecast the answer to better food security? To better early warning? To better famine prevention? *Internet J. Afr. Stud.* (2), available on-line, 1997.

Thomas, D. A. G., and D. Sporton, Understanding the dynamics of social and environmental variability, *Appl. Geogr.*, *17*(1), 11–27, 1997a.

Thomas, D. A. G.. and D. Sporton *Environmental Change and Poverty in Kalahari Pastoral Systems: Full Report of Research Activities and Results*, available on-line, http://www.shef.ac.uk/uni/academic/I-M/idry/Esrcreport.html, 1997b.

United Nations Department of Humanitarian Affairs (UNDHA), *Glossary: Internationally Agreed Glossary of Basic Terms Related to Disaster Management*, Geneva, UNDHA, 1992.

Wolde-Georgis, T., El Niño and drought early warning in Ethiopia, *Internet J. Afr. Stud.* (2), available on-line, 1997.

Yath, Y. A., Dinka migrants in Khartoum: Coping strategies in the face of economic and political hardships. The example of the Suq El Markazi squatter settlement, Sudan, in H. G. Bohle, T. E. Downing, J. O. Field, and F. M. Ibrahim (Eds.), *Coping with Vulnerability and Criticality: Case Studies on Food-Insecure People and Places*, Saarbrucken, Fort Lauderdale, Breitenbach, 1993, pp. 59–80.

CHAPTER 39

DROUGHT IN THE U.S. GREAT PLAINS

DONALD A. WILHITE

1 INTRODUCTION

Drought, a normal feature of the climate for virtually all portions of the United States, is one of the defining characteristics of the Great Plains region. Early maps referred to this region as the Great American Desert, a belief attributed to the explorations of Zebulon Pike in the early 1800s (Brown, 1948). The region's past is firmly rooted in the drought of the 1890s and, in particular, the Dust Bowl years of the 1930s (Hurt, 1981). More recently, droughts have occurred at regular intervals, affecting all portions of a region that stretches from Texas and New Mexico northward through the Dakotas and Montana into the Prairie Provinces of Alberta, Saskatchewan, and Manitoba. In reality, it is rare for drought not to occur in the region each year, a fact that has forced considerable adjustments from a predominantly agricultural economy. Irrigation development and many other technological adjustments in the post-1930s era have improved the resilience of the region to the ravages of drought, but drought continues to produce devastating and widespread impacts.

The purpose of this chapter is to discuss drought in the context of the Great Plains. In that process, I will review some of the basic concepts of drought. A grasp of these concepts is essential to understanding the history and impacts of drought in the Great Plains and the region's continuing vulnerability to this insidious natural hazard. Current and future attempts to lessen drought impacts in the region are also discussed.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

2 CONCEPT OF DROUGHT: DEFINITION AND TYPES

Drought is the consequence of a natural reduction in the amount of precipitation received over an extended period of time, usually a season or more in length, although other climatic factors (such as high temperatures, high winds, and low relative humidity) are often associated with it in many regions of the world and can significantly aggravate the severity of the event (Wilhite, 1992, 2000). High winds and low relative humidity aggravated the effects of the drought of the 1930s in the Great Plains. Drought is also related to the timing (i.e., principal season of occurrence, delays in the start of the rainy season, occurrence of rains in relation to principal crop growth stages) and the effectiveness (i.e., rainfall intensity, number of rainfall events) of the rains. Thus, each drought is unique in its climatic characteristics and impacts. Likewise, society is changing in response to increasing and shifting population, new technologies, government policies, and social behavior. These factors alter vulnerability, a fact that will be discussed in greater detail later in this chapter.

Drought is a temporary aberration that occurs in high- and low-rainfall regions (Wilhite, 1992). Although droughts are commonly associated with the Great Plains and other semiarid regions, it is difficult for many people to visualize drought occurring in more humid regions such as the eastern United States, Southeast Asia, Brazil, or western Europe. This fact emphasizes both the regional and relative nature of drought.

Drought differs from other natural hazards (e.g., floods, tropical cyclones, and earthquakes) in several ways. First, drought is a slow-onset natural hazard. It is often referred to as a creeping phenomenon (Tannehill, 1947). The effects of drought accumulate slowly over a considerable period of time and may linger for years after the termination of the event. As a result, the onset and end of drought are difficult to determine. Even today, with more sophisticated monitoring technology, climatologists struggle to recognize the onset of drought, and scientists and policymakers continue to debate the basis (i.e., criteria) for declaring an end to a drought. Second, the absence of a precise and universally accepted definition of drought adds to the confusion about whether or not a drought exists and, if it does, its degree of severity. Realistically, definitions of drought must be region-specific and application (or impact) specific. Wilhite and Glantz (1985) analyzed more than 150 definitions in their classification study, and many more definitions exist. Although the definitions are numerous, many do not adequately define drought in meaningful terms for scientists or policymakers. Third, drought impacts are nonstructural and spread over a larger geographical area than are damages that result from other natural hazards. For example, a recent analysis of drought occurrence by the U.S. National Drought Mitigation Center for the 48 contiguous states in the United States demonstrated that severe and extreme drought affected more than 25% of the country in 27 of the past 100 years.

Because drought affects virtually all regions of the world and many economic and social sectors, scores of definitions exist. Impacts are complex, vary on spatial and temporal scales, and depend on the societal context of drought. The impacts of drought in the Great Plains will differ from those experienced in the southeast,

northeast, or far western portions of the United States. As a result, it is not possible to formulate a definition of drought that is universally acceptable in each of these settings. Wilhite and Glantz (1985) concluded that definitions of drought should reflect a regional bias since water supply is largely a function of climatic regime.

Drought has been grouped by type as follows: meteorological, agricultural, hydrological, and socioeconomic (Wilhite and Glantz, 1985). Meteorological (or climatological) drought is expressed solely on the basis of the degree of dryness (often in comparison to some normal or average amount) and the duration of the dry period. The *Encyclopedia of Climate and Weather* (Schneider, 1996) defines drought as an extended period—a season, a year, or several years—of deficient rainfall relative to the statistical multiyear mean for a region. This definition identifies two critical components that must be accounted for in a viable definition—intensity and duration. Meteorological drought definitions must be considered as region specific since the atmospheric conditions that result in deficiencies of precipitation are climate regime dependent. In the Great Plains, the distribution of precipitation is seasonal: Approximately 70% of the precipitation in this region occurs during the 6-month period from April to September. Definitions that differentiate meteorological drought on the basis of the number of days with precipitation less than some specified threshold (e.g., for Britain, 15 days, none of which received as much as 0.25 mm of precipitation; British Rainfall Organization, 1936) rather than the magnitude of the deficiency over some period of time would be inappropriate for the Great Plains.

Agricultural drought links various characteristics of meteorological drought to agricultural impacts, focusing on precipitation shortages, differences between actual and potential evapotranspiration (ET), soil water deficits, and so forth. Rosenberg (1980) defined agricultural drought as a climatic excursion involving a shortage of precipitation sufficient to adversely affect crop production or range productivity. A definition of agricultural drought should account for the variable susceptibility of crops at different stages of crop development. The impacts of drought are crop specific because the most weather-sensitive phenological stages vary between crops. Planting dates and maturation periods also vary between crops and locations. A period of high-temperature stress that occurs in association with dry conditions may coincide with a critical weather-sensitive growth stage for one crop while missing a critical stage for another crop. Agricultural planning can often reduce the risk of drought impact on crops by altering the crop, genotype, planting date, and cultivation practices. Considerable progress has been made in the Great Plains and elsewhere in applying various types of adaptive strategies to reducing the impacts of drought on crop and rangeland (Rosenberg, 1980, 1986).

Agricultural droughts usually take 3 months or more to develop, but this time period can vary considerably, depending on the timing of the initiation of the precipitation deficiency. For example, in the Great Plains a significant dry period during the winter season may have few, if any, impacts for many locales. However, if this deficiency continues into the growing season, the impacts may magnify quickly since low precipitation during the autumn and winter season results in low soil moisture recharge rates, leading to deficient soil moisture at spring planting. Although the region's agriculture is usually the first to feel the effects of drought,

a prolonged dry period can result in significant disruptions in other sectors, particularly water-based transportation, energy production, municipal water supplies, and recreation-based businesses. Also, a short-lived drought of less than 3 months can have serious impacts on crop yields if it occurs during the critical crop growth stages and is accompanied by high temperatures.

Hydrological droughts are associated more with the effects of periods of precipitation shortfall on surface or subsurface water supply (i.e., streamflow, reservoir and lake levels, groundwater) than with precipitation shortfalls directly (Dracup et al., 1980; Klemês, 1987). Hydrological droughts are usually out of phase or lag the occurrence of meteorological and agricultural droughts. Meteorological droughts result from precipitation deficiencies; agricultural droughts are largely the result of soil moisture deficiencies. More time elapses before precipitation deficiencies are detected in other components of the hydrological system (e.g., reservoirs, groundwater). As a result, impacts from hydrological drought are out of phase with those in other economic sectors. Water in hydrological storage systems (e.g., reservoirs, rivers) is often used for multiple and competing purposes (e.g., power generation, flood control, irrigation, recreation), further complicating the sequence and quantification of impacts. Competition for water in these storage systems escalates during drought, and conflicts between water users increase significantly. For example, changing water-use patterns and the series of drought years that occurred in the Missouri River basin between 1987 and 1992 resulted in significant conflicts between upstream and downstream water users.

Finally, socioeconomic drought associates the supply and demand of some economic good or service with elements of meteorological, hydrological, and agricultural drought. For example, the supply of some economic good (e.g., water, hay, hydroelectric power) is highly sensitive to the vagaries of weather. In most instances, the demand for that good is increasing as a result of increasing population and/or per capita consumption. Therefore, drought could be defined as occurring when the demand for that good exceeds supply as a result of a weather-related supply shortfall (Sandford, 1979). This concept of drought supports the strong symbiosis that exists between drought and human activities. Thus, the incidence of drought could increase because of a change in the frequency of the physical event, a change in societal vulnerability to water shortages, or both. For example, poor land-use practices such as overgrazing can decrease animal carrying capacity and increase soil erosion, which exacerbates the impacts of and vulnerability to future droughts. Overdrafts of groundwater, such as have been occurring in the southern portions of the Ogallala aquifer, will affect future vulnerability to drought in the Great Plains since access to groundwater is a primary adaptive strategy employed to alleviate the effects of drought.

3 DROUGHT CLIMATOLOGY OF THE GREAT PLAINS

Historical climate records for the Great Plains provide only a brief snapshot of the drought climatology for the region. For most portions of the region, climatic records

cover the period since about 1900, and at only a few locations. To learn more about the occurrence and patterns of drought before that time, tree-ring data can be used to reconstruct the drought history of the region. These data provide insights into past climates extending back many centuries. Figure 1 is adapted from the work of Weakly (1965) for western Nebraska. His work was based largely on tree rings from Red Cedar for a 748-year period, from 1210 to 1958. The results of his study showed the occurrence of 21 drought periods of 5 years or more duration. The most remarkable of these drought periods was from 1276 to 1313, a period of 38 years. However, the average duration of the droughts was 12.8 years. A similar study was conducted in northern and southern Texas by Stahle and Cleaveland (1988) for the period 1698 to 1980. This analysis revealed numerous drought events during this nearly 300-year study period. The most severe individual drought years before 1900 for north Texas were 1772, 1790, 1805, 1855, 1872, and 1887. The years 1917, 1925, 1939, and 1956 were the driest years of the twentieth century. For south Texas, individual drought years before 1900 were 1790, 1805, 1855, 1857, and 1887, and the driest years of the twentieth century occurred in 1917, 1925, 1956, 1967, and 1971. Other tree-ring studies in the region confirm the recurrent nature of single and multiyear droughts in the region (Stockton and Meko, 1975, 1983).

Figure 2 provides a historical perspective of the percent area of the Great Plains in severe to extreme drought, according to the Palmer Drought Severity Index (PDSI) (Palmer, 1965) from 1895 to 1995. The PDSI is a meteorological drought index that integrates many variables in a water balance accounting procedure. This index is calculated routinely for each of the climate divisions in the United States. PDSI values commonly range from + 4.0 (extreme wetness) to − 4.0 (extreme drought), although values above and below these levels are often computed. For example, during August 1977, PDSI values reached − 7.0 in parts of the upper Midwest. For the Great Plains region, Figure 2 illustrates three interesting characteristics. First, the percent area in drought is highly variable from year to year. The peak drought year was 1934 in which 95% of the region was experiencing severe to extreme drought; other severe drought years were 1936 and 1956, with 90 and 80%, respec-

Figure 1 Periods of drought in western Nebraska, five or more years in duration, 1200–1960. Periods of drought shown in black. Numbers in parentheses following year indicate length of drought period. Average duration of drought: 12.8 years (adapted from Weakly, 1965).

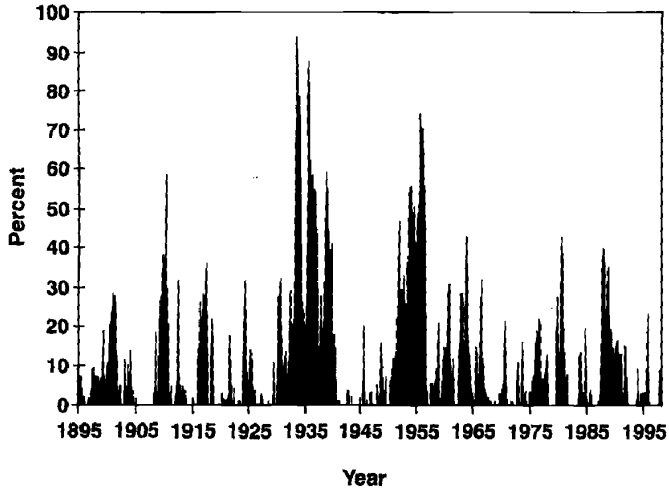


Figure 2 Percent area of the Great Plains experiencing severe to extreme drought, 1895–1995.

tively, of the region in severe and extreme drought. Second, it is rare for severe drought not to be occurring at some location in the region in every single year during 1895 to 1995. Third, clusters of drought years, while rare, are particularly noticeable in the 1930s, mid-1950s, late 1970s, and late 1980s to early 1990s. Multiyear droughts are important because impacts magnify as drought continues into a second and subsequent years. As surface and subsurface water supplies are gradually depleted, more economic sectors are affected. It often takes years for these systems (i.e., reservoirs, groundwater) to recover following an extended drought episode.

Figure 3 illustrates the percent area of three river basins in the Great Plains region in severe to extreme drought during the period 1895 to 1995. The river basins shown are the Missouri, Arkansas–White–Red, and Rio Grande. These three basins encompass most of the area of the Great Plains. The drought climatology of these three basins displays characteristics similar to those shown in Figure 2. However, there are two important differences. First, largely because of the spatial characteristics of drought, it is more common for drought to affect nearly 100% of these basins. This is especially noticeable for the Arkansas–White–Red and Rio Grande basins. This information is important to planners and water supply managers. Second, the pattern of drought differs from one portion of the region to another. For example, the 1930s drought was of much greater duration in the Missouri Basin than in the Rio Grande. By contrast, the 1950s drought was of greater duration and intensity in the Rio Grande basin. This illustrates the regional nature of drought and the fact that planning must be based on the drought of record for the area of interest. Tree-ring data can help scientists reconstruct the long-term climatic history of the region.

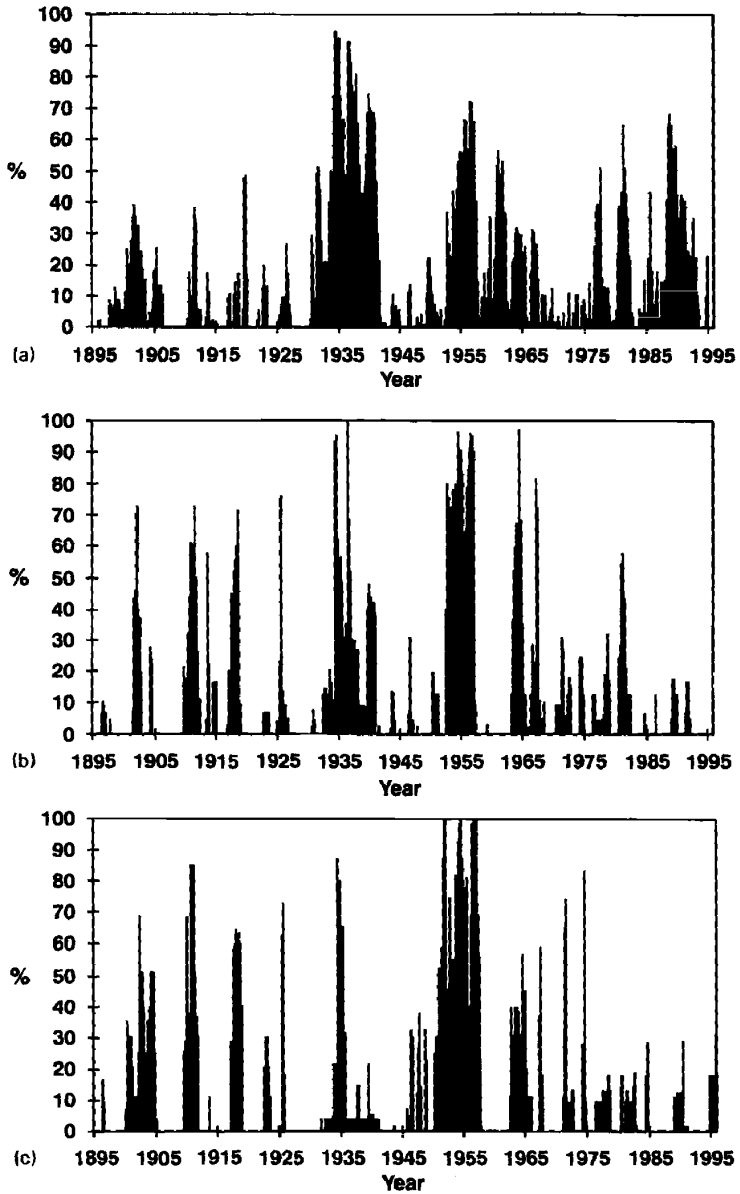


Figure 3 Percent area of (a) Missouri Basin, (b) Arkansas–White–Red Basin, and (c) Rio Grande Basin in severe and extreme drought (i.e., ≤ -3.0) during the period 1895–1995.

4 IMPACTS OF DROUGHT

Drought produces a complex web of impacts that not only ripple through many sectors of the economy but may be experienced well outside the affected region, extending even to the global scale. This complexity is largely caused by the dependence of so many sectors on water for producing goods and providing services. Agricultural production in the Great Plains is of critical importance to food production in the United States. Substantial drought-related production losses not only affect food supplies and prices in this country but also have serious implications for the many nations that depend on U.S. grain exports to offset domestic supply shortfalls.

Impacts from drought are commonly classified as direct or indirect. Reduced crop, rangeland, and forest productivity; increased fire hazard; reduced water levels; increased livestock and wildlife mortality rates; and damage to wildlife and fish habitat are a few examples of direct impacts. The consequences of these impacts illustrate indirect impacts. For example, a reduction in crop, rangeland, and forest productivity may result in reduced income for farmers and agribusiness, increased prices for food and timber, unemployment, reduced government tax revenues because of decreased expenditures, increased crime, foreclosures on bank loans to farmers and businesses, migration, and disaster relief programs. Direct or primary impacts are usually of a biophysical nature. Conceptually speaking, the more removed the impact from the cause, the more complex the link to the cause.

Because of the number of affected groups and sectors associated with drought, the geographic size of the area affected, and the difficulties in quantifying environmental damages and personal hardships, the precise determination of the financial costs of drought is a formidable challenge. The economic costs and losses associated with drought are highly variable from year to year. These costs and losses are also quite variable from one drought year to another in the same place, depending on timing, intensity, and spatial extent of the droughts.

The impacts of drought are commonly classified as economic, environmental, and social. Table 1 presents a comprehensive list of the impacts associated with drought. This list represents the experiences of the Great Plains and many other drought-prone areas of the world. Although drought produces impacts that are regionally distinct, there are many similarities in the types of impacts experienced from one region to another. Many economic impacts occur in broad agricultural and agriculturally related sectors, including forestry and fisheries, because of the reliance of these sectors on surface and subsurface water supplies. In addition to obvious losses in yields in both crop and livestock production, drought is associated with increases in insect infestations, plant disease, and wind erosion. Droughts also bring increased problems with insects and diseases to forests and reduce growth. The incidence of forest and range fires increases substantially during extended droughts, which in turn places both human and wildlife populations at higher levels of risk.

Income loss is another indicator used in assessing the impacts of drought because so many sectors are affected. Reduced income for farmers has a ripple effect, as their ability to purchase goods and services is limited. Thus, many retailers experience

TABLE 1 Classification of Drought-Related Impacts (Costs and Losses)

Problem Sectors	Impacts
Economic	Loss from crop production
	Annual and perennial crop losses; damage to crop quality
	Reduced productivity of cropland (wind erosion, etc.)
	Insect infestation
	Plant disease
	Wildlife damage to crops
	Loss from dairy and livestock production
	Reduced productivity of rangeland
	Forced reduction of foundation stock
	Closure/limitation of public lands to grazing
	High cost/unavailability of water for livestock
	High cost/unavailability of feed for livestock
	High livestock mortality rates
	Increased predation
	Range fires
	Loss from timber production
	Forest fires
	Tree disease
	Inset infestation
	Impaired productivity of forest land
	Loss from fishery production
	Damage to fish habitat
	Loss of young fish due to decreased flows
	Loss of national economic growth, retardation of economic development
	Income loss for farmers and others directly affected
	Loss of farmers through bankruptcy
	Loss to recreational and tourism industry
	Loss to manufacturers and sellers of recreational equipment
	Increased energy demand and reduced supply because of drought-related power curtailments
	Costs to energy industry and consumers associated with substituting more expensive fuels (oil) for hydroelectric power
	Loss to industries directly dependent on agricultural production (e.g., machinery and fertilizer manufacturers, food processors, etc.)
	Decline in food production/disrupted food supply
	Increase in food prices
	Increased importation of food (higher costs)
	Disruption of water supplies
	Unemployment from drought-related production declines
	Strain on financial institutions (foreclosures, greater credit risk, capital shortfalls, etc.)
	Revenue losses to federal, state, and local governments (from reduced tax base)
	Deters capital investment, expansion
	Dislocation of businesses

(*Continued*)

TABLE 1 ***(continued)***

Problem Sectors	Impacts
	Revenues to water supply firms
	Revenue shortfalls
	Windfall profits
	Loss from impaired navigability of streams, rivers and canals
	Cost of water transport or transfer
	Cost of new or supplemental water resource development
Environmental	Damage to animal species
	Reduction and degradation of fish and wildlife habitat
	Lack of feed and drinking water
	Disease
	Increased vulnerability to predation (e.g., from species concentration near water)
	Loss of biodiversity
	Wind and water erosion of soils
	Reservoir and lake drawdown
	Damage to plant species
	Water quality effects (e.g., salt concentration, increased water temperature, pH, dissolved oxygen)
	Air quality effects (dust, pollutants)
	Visual and landscape quality (dust, vegetative cover, etc.)
	Increased fire hazard
	Estuarine impacts; changes in salinity levels, reduced flushing
Social	Increased groundwater depletion (mining), land subsidence
	Loss of wetlands
	Loss of cultural sites
	Insect infestation
	Food shortages (decreased nutritional level, malnutrition, famine)
	Loss of human life (e.g., food shortages, heat)
	Public safety from forest and range fires
	Conflicts between water users, public policy conflicts
	Increased anxiety
	Loss of aesthetic values
	Health-related low flow problems (e.g., diminished sewage flows, increased pollutant concentrations, etc.)
	Recognition of institutional constraints on water use
	Inequality in the distribution of drought impacts/relief
	Decreased quality of life in rural areas
	Increased poverty
	Reduced quality of life, changes in life-style
	Social unrest, civil strife
	Population migration (rural to urban areas)
	Reevaluation of social values
	Increased data/information needs, coordination of dissemination activities
	Loss of confidence in government officials
	Recreational impacts

[a]Input from working groups was used to modify a table from Wilhite (1992).

significant reductions in sales. This leads to unemployment, increased credit risk for financial institutions, capital shortfalls, and loss of tax revenue for local, state, and federal government. The recreation and tourism industries are also affected because of less discretionary income. Prices for food, energy, and other products increase as supplies are reduced. In some cases, local supply shortfalls for certain goods will result in the importation of these goods from outside the stricken region. Reduced water supply impairs the navigability of rivers and results in increased transportation costs because products must be transported by rail or truck. Hydropower production is also significantly reduced. For example, hydropower generation was 25 to 40% below average for large sections of the country in 1988, resulting in serious revenue losses for the industry (Wilhite, 1993a).

Environmental losses are the result of damages to plant and animal species, wildlife habitat, and air and water quality; forest and range fires; degradation of landscape quality; loss of biodiversity; and soil erosion. Some of the effects are short term and conditions quickly return to normal following the end of the drought. Other environmental effects linger for some time or may even become permanent. Wildlife habitat, for example, may be degraded through the loss of wetlands, lakes, and vegetation. However, many species will eventually recover from this temporary aberration. The degradation of landscape quality, including increased soil erosion, may lead to a more permanent loss of biological productivity of the landscape. Although environmental losses are difficult to quantify, growing public awareness and concern for environmental quality has forced public officials to focus greater attention and resources on these effects.

Social impacts mainly involve public safety, health, conflicts between water users, reduced quality of life, and inequities in the distribution of impacts and disaster relief. Many of the impacts specified as economic and environmental have social components as well. Population out-migration was a significant problem in the Great Plains in response to the 1930s drought and continues to be a major problem in many countries.

As with all natural hazards, the economic impacts of drought are highly variable within and between economic sectors and geographic regions, producing a complex assortment of winners and losers with the occurrence of each disaster. For example, decreases in agricultural production result in enormous negative financial impacts on farmers in drought-affected areas, at times leading to foreclosure. This decreased production also leads to higher grain, vegetable, and fruit prices. These price increases have a negative impact on all consumers as food prices increase. However, farmers outside the drought-affected area with normal or above-normal production or those with significant grain in storage reap the benefits of these higher prices. Similar examples of winners and losers could be given for other economic sectors as well.

5 DROUGHT MANAGEMENT

Although drought is a natural hazard, the term *drought management* implies that human intervention can reduce vulnerability and impacts (Wilhite, 1993b). To be

successful in this endeavor, many disciplines must work together in tackling the complex issues associated with detecting, responding to, and preparing for the inevitability of future events. To improve our management of drought requires that we view drought as having both a natural component and a social component. In other words, the risk associated with drought in the Great Plains or any region is a product of both its exposure to the event (i.e., probability of occurrence at various severity levels) and the vulnerability of society to the event. The natural event (i.e., meteorological drought) is a result of the occurrence of persistent large-scale disruptions in the global circulation pattern of the atmosphere. Exposure to drought varies spatially and there is little, if anything, that we can do to alter the occurrence of meteorological drought. As previously discussed, the Great Plains has historically had a very high incidence of drought. Certainly, there is no reason to believe that this incidence will diminish in the future. In fact, output from general circulation models suggest that the interiors of midlatitude continents are likely to become drier as a result of increasing concentrations of greenhouse gases in the atmosphere. This would result not only from increasing temperatures and associated increases in evapotranspiration but also from possible changes in the amount, seasonal distribution, and effectiveness of precipitation. The result may be a net loss in soil moisture during the growing season. If these projections are correct, the incidence of drought in the Great Plains may increase.

Vulnerability, on the other hand, is determined by factors such as population, demographic characteristics, technology, policy, and social behavior. These factors change over time, and thus vulnerability is likely to increase or decrease in response to these changes. Subsequent droughts in the same region will have different effects, even if they are identical in intensity, duration, and spatial characteristics, because societal characteristics have changed.

Much has been done to lessen societal vulnerability to drought in the Great Plains. The widespread adoption of irrigation, conservation tillage practices, soil evaporation reduction, snow management, and irrigation scheduling have all proved effective in stabilizing agricultural production in a region exposed to the vagaries of weather. However, we must continue to implement new mitigation techniques and preparedness strategies in the face of drought. Recent droughts illustrate our continuing vulnerability to extended periods of water shortage. In 1988, drought affected nearly 40% of the nation and resulted in nearly $16 billion in agricultural losses (Riebsame, et al. 1991). In the Great Plains, this drought had serious impacts on spring wheat yields, reducing yields by 54% (Riebsame, et al. 1991). In 1989, winter wheat and sorghum yields were reduced substantially in parts of the central Great Plains. In 1996, drought in the Southwest and southern Great Plains states resulted in substantive agricultural losses, increased incidence of forest and range fires, municipal water supply problems, and losses in recreation and tourism. In Texas alone, the 1996 drought losses were estimated to be $6.5 billion (Boyd, 1996). The Federal Emergency Management Agency (FEMA, 1995) recently estimated annual losses resulting from drought in the United States at $6 to $8 billion.

Drought planning is one of the mechanisms being employed by many states in the Great Plains and nationwide to reduce the economic losses and personal hardships

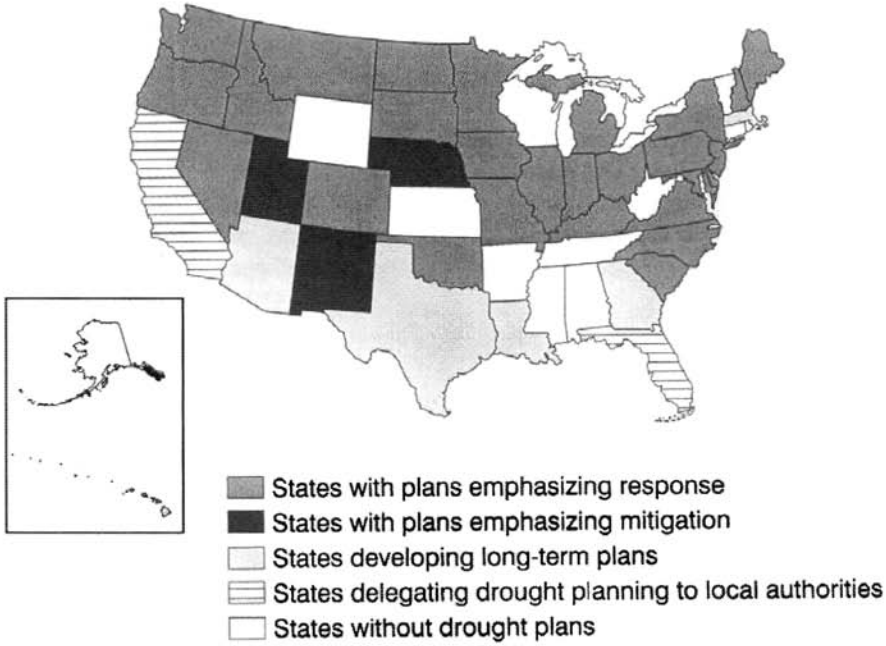


Figure 4 Status of state drought plans as of January 2001.

associated with drought. The number of states in the United States with drought plans has grown from 3 in 1982 to 27 in 1997 (Fig. 4) (Wilhite, 1997a). In addition to the states that now have plans, six states (Alabama, Louisiana, Texas, New Mexico, Arizona, and Pennsylvania) are at various stages of plan development or have expressed intent to develop a plan. In the U.S. portion of the Great Plains region, all states except Kansas and Wyoming have developed plans. Alberta has also undertaken some initial steps in drought plan development.

The basic goal of state drought plans is to improve the effectiveness of state response and preparedness efforts. This is accomplished by improving monitoring and early warning, impact and vulnerability assessment, and preparedness, response, recovery, and mitigation programs (Wilhite, 1997b). These plans are also directed at improving coordination and building partnerships within agencies of state government and between state, local, and federal governments. The growth in the number of states with drought plans suggests an increased concern about the potential impact of extended water shortages and an attempt to address those concerns through planning. However, more attention needs to be placed on mitigation, defined as short- and long-term actions, programs, or policies implemented in advance that reduce the degree of risk to people, property, and productive capacity (Wilhite, 1997b).

In 1997, the Western Drought Coordination Council (WDCC) was formed under the auspices of the Western Governors' Association (WGA) as a result of a memor-

andum of understanding between WGA and key federal agencies (Departments of Agriculture, Interior, and Commerce, FEMA, and the Small Business Administration). This activity represents an important attempt to build regional partnerships between local, state, federal, and tribal governments to reduce the impacts of future drought events in the western states through greater attention to planning and mitigation. The Great Plains states are actively participating in the WDCC. These types of improved institutional arrangements, in combination with the existence of drought plans and the application of new technologies, are an important new trend in mitigating the effects of drought in the Great Plains states and elsewhere (Wilhite, 1997c).

6 SUMMARY

Drought is a complex, recurrent, and insidious natural hazard that has historically resulted in significant impacts in the Great Plains. Its impacts are far-reaching and may linger for months or even years beyond the termination of the event. The economic, social, and environmental impacts of drought result from complex interactions between physical and social systems, and they are difficult to quantify. Scientists and policymakers must understand the characteristics of drought and appreciate the magnitude and complexity of impacts in order for viable assessment and response strategies to be established. The aim of these strategies is to reduce societal vulnerability to periods of water shortages.

Drought inflicts considerable pain and hardship on society. The impacts of contemporary droughts in the Great Plains have demonstrated this fact repeatedly over the past several decades. Drought illustrates, in innumerable ways, the vulnerability of economic, social, political, and environmental systems to a variable climate. It also illustrates the dependencies that exist between systems, reinforcing the need for improved coordination within and between levels of government.

Extended periods of normal or benign weather conceal the vulnerability of societies to climate variability, but drought exposes these sensitivities. Projected changes in climate because of increased concentrations of carbon dioxide and other atmospheric trace gases suggest a possible increase in the frequency and intensity of severe drought in the Great Plains region. In a region where the incidence of drought is already high, any increase in drought frequency will place even greater pressure on the region's already limited water supplies. It is critical for us to assess our exposure and vulnerabilities to drought and take the actions necessary to reduce risk through enhanced mitigation and preparedness.

REFERENCES

Boyd, J., *Southwest Farmers Battle Record Drought*, United Press International, May 30, 1996.

British Rainfall Organization, *British Rainfall*, Air Ministry, Meteorological Office, cited in World Meteorological Organization, *Drought and Agriculture*, Technical Note 138, Geneva, Switzerland, 1936.

Brown, R. H., *Historical Geography of the United States*, Harcourt, Brace, and World, New York, 1948.

Dracup, J. A., K. S. Lee, and E. G. Paulson, Jr., On the definition of droughts, *Water Resourc. Res.*, *16*(2), 297–302, 1980.

Federal Emergency Management Agency (FEMA), *National Mitigation Strategy: Partnerships for Building Safer Communities*, FEMA, Washington, DC, 1995.

Hurt, R. D., *An Agricultural and Social History: The Dust Bowl*, Nelson-Hall, Chicago, IL, 1981.

Klemês, V., Drought prediction: A hydrological perspective, in D. A. Wilhite and W. E. Easterling (Eds.), *Planning for Drought: Toward a Reduction of Societal Vulnerability*, Westview, Boulder, CO, 1987, Chapter 7.

Palmer, W. D., *Meteorological Drought*, Research Paper No. 45, U.S. Weather Bureau, Washington, DC, 1965.

Riebsame, W. E., S. A. Changnon, and T. R. Karl, *Drought and Natural Resources Management in the United States*, Westview, Boulder, CO, 1991.

Rosenberg, N. J. (Ed.), *Drought in the Great Plains: Research on Impacts and Strategies*, Water Resources Publications, Littleton, CO, 1980.

Rosenberg, N. J., Adaptations to adversity: Agriculture, climate and the Great Plains of North America, *Great Plains Q.*, *6*, 202–217, 1986.

Sandford, S., Towards a definition of drought, in M. T. Hinchey, (Ed.), *Botswana Drought Symposium*, Botswana Society, Gaborone, Botswana, 1979.

Schneider, S. H. (Ed.), *Encyclopedia of Climate and Weather*, Oxford University Press, New York, 1996.

Stahle, D. W., and M. K. Cleaveland, Texas drought history reconstructed and analyzed from 1698 to 1980, *J. Climate*, *1*, 59–74, 1988.

Stockton, C. W., and D. M. Meko, A long-term history of drought occurrence in the western United States as inferred from tree rings, *Weatherwise*, December 1975. pp. 245–249.

Stockton, C. W., and D. M. Meko, Drought recurrence in the Great Plains as reconstructed from long-term tree-ring records, Journal of Climate and Applied Meteorology. vol. 22, pp. 17–29.

Tannehill, I. R., *Drought: Its Causes and Effects*, Princeton University Press, Princeton, NJ, 1947.

Weakly, H. D., Recurrence of drought in the Great Plains during the last 700 years, *Agric. Eng.*, February 1965, Vol. 46, p. 85.

Wilhite, D. A., Drought, in *Encyclopedia of Earth System Science*, Vol. 2, Academic, San Diego, CA, 1992, pp. 81–92.

Wilhite, D. A., Understanding the phenomenon of drought, *Hydro-Review*, *12*, 136–148, 1993a.

Wilhite, D. A. (Ed.), *Drought Assessment, Management, and Planning: Theory and Case Studies*, Kluwer Academic, Boston, MA, 1993b.

Wilhite, D. A., State actions to mitigate drought: Lessons learned, *J. Am. Water Res. Assoc.*, *33*(5): 961–968, 1997a.

Wilhite, D. A., Improving drought management in the West, Report to the Western Water Policy Review Advisory Commission, 1997b.

Wilhite, D. A., Responding to drought: Common threads from the past, visions for the future, *J. Am. Water Res. Assoc.*, *33*(5): 951–959, 1997c.

Wilhite, D. A. and M. H. Glantz, Understanding the drought phenomenon: The role of definitions, *Water Int.*, *10*, 111–120, 1985.

Wilhite, D. A. (editor), 2000, Draught: A Global Assessment. Routledge Publishers. London, England.

CHAPTER 40

FLOODS ON THE MISSISSIPPI RIVER SYSTEM OF THE UNITED STATES

STANLEY A. CHANGNON

1 INTRODUCTION

For the past 150 years a titanic struggle has been underway between the human occupants of the Mississippi River system and its floods. A flood of some type occurs somewhere in this giant basin each year, and every 2 to 10 years a massive flood encompasses a fourth or more of the 3.2 million km^2 basin. Humans first tried to control the floods with structures: channel straightening, levees, dams, and reservoirs; but after 100 years and the expenditure of billions of dollars, losses to property and lives continued to grow. Efforts since the 1950s to encourage land-use changes in flood-prone areas and the use of flood insurance often have been thwarted by government relief payments to flood victims, plus a continuing human desire to ignore the threat and reside in floodplains. Only 10% of the residents of flooded areas in the massive floods on the Mississippi and Missouri Rivers in 1993 and on the Ohio River in 1997 had flood insurance. Ever growing urban sprawl in the floodplains, intense use of the rivers for shipping, dense surface transportation networks in the floodplains, and major cities and industrial complexes built along major rivers help keep the basin highly vulnerable to today's floods.

The river system brings enormous economic value to the United States, but interests in protecting and enhancing the natural environment of the river are commonly in direct conflict with economic interests, and flood mitigation is often caught in the middle of the debate about how to manage the river system to satisfy all interests. The massive 1993 flood losses and responses, costing $18 billion, brought about needed changes in crop and flood insurance. The huge and costly

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

infrastructure built to control the major rivers for floods and navigation is aging and is beginning to need replacement. This offers an opportunity for more correctly handling flood mitigation to satisfy the complex mix of economic, human, and environmental interests (Shabman, 1994). However, if the past is a predictor of the future, this seems unlikely, but we do know one thing about the future—major record-setting floods will continue to occur.

Thirteen lessons emanating from the recent struggles between humans and floods include:

1. After massive expenditures to control flooding, flooding in the Mississippi River system is still the most costly natural hazard of the region.
2. Floods occurring at various scales, from localized flash floods to enormous basin-scale floods cannot be controlled, although existing control works help to reduce losses.
3. Forecasting and warning of floods have improved, leading to a reduction in lives lost, but people still are killed due to failure to understand the risks or to receive warnings.
4. Surface and river transportation systems suffer major damages and costly delays.
5. Communities suffer costly damages to water and sewage treatment plants and huge costs for postflood cleanup, and many flood-prone communities consciously chose to not construct adequate flood protection systems.
6. Flooding in suburban areas is increasing, due to construction in questionable areas, a lack of control on residential and commercial developments with inadequate stormwater control systems, and a lack of regional plans for managing floodwaters.
7. Millions of individuals continue to risk flood losses by failing to purchase flood insurance, relying on government relief, and simply not appreciating their risk.
8. Environmental impacts of floods are mixed—the worst relate to soil erosion and river pollution by chemicals—but flooding generally helps floodplain ecosystems.
9. Agricultural impacts from floods can be enormous, particularly if the floods occur during the growing season.
10. Government relief for flood victims and communities remains a major response that often acts to discourage doing the right things in floodplains.
11. Rivers will reclaim their floodplains in extreme floods, and it is wise to work with the river and not against it in deciding where and how to build levees and in rebuilding other control structures such as the aging lock-and-dam system.
12. Future use of the floodplains of the Mississippi River system will require a careful balance between controlling for natural variations for navigation and flood protection purposes or for benefits to the ecosystem.

13. In sum, the nation's policy philosophy about flood mitigation must change: The nation needs to move beyond reliance on political responses and solutions to inappropriate uses of floodplains. Individuals must assume responsibility for their locational decisions, not the government, and future government policies must stand firm over time.

To understand the floods of this huge river system, and the impacts the floods create, requires background information about the physical and human setting of the river system. This setting has its roots in the basin's physical formation and configuration and the history of settlement and ensuing development of the basin. Today's impacts of flooding integrate many decades of massive, costly efforts to mitigate flooding in the Mississippi River system.

Physical Setting

The flow of the Mississippi River ranks as the world's third largest behind the Amazon and Congo. The outflow of the river into the Gulf of Mexico averages 173,600 m^3/s and represents 5% of all the freshwater discharged into the oceans of the world (Tarbuck and Lutgens, 1984). The variability of precipitation falling in the basin is large, reflected in a 1-year record low flow of 75,000 m^3/s, and a record 1-year high of 600,000 m^3/s, more than three times the long-term average. The basin occupies all or part of 34 states and 41% of the contiguous United States, sprawling over 3.2 million km^2. Headwaters exist in the Rocky Mountains of the west, the Appalachian Mountains of the east, and the forests of the northern United States. (Fig. 1).

The enormous basin embraces five major climatic regions including humid continental, semiarid steppe, and wet subtropical climates. Four very different physiographic regions are found in the basin including the world's most productive soils and seven very different vegetative regions (White et al., 1979). This diversification in the physical setting greatly affects the type and frequency of floods that occur. Millions of years ago, the river's delta began forming near Cairo, Illinois (Fig. 1), and subsequently advanced 1600 km to the south. New Orleans, the river's major port city, rests where ocean waters existed only 5000 years ago. The enormity of the basin's erosion and sediment transport of the river is shown by the fact that the river deposits 750 million tons of sand, silt, and clay annually into the Gulf of Mexico (Tarbuck and Lutgens, 1984). The sediment load, resulting from the record 1993 flood, led to a discharge of 2.1 billion tons (Bhowmik, 1996).

The basin is so large that it embraces two other enormous river basins—those of the Ohio and Missouri Rivers, and the Mississippi alone is so huge that it is divided hydrologically at St. Louis into the upper and lower basins. Due to climatic differences and the basin's enormity, there has never been a flood that encompassed the entire Mississippi basin. The net effect of the basin's physical situation (size and climatic factors creating floods) is that floods only occur at a given time on one, or infrequently two, of these large basins like the Ohio, the Upper Mississippi, or the

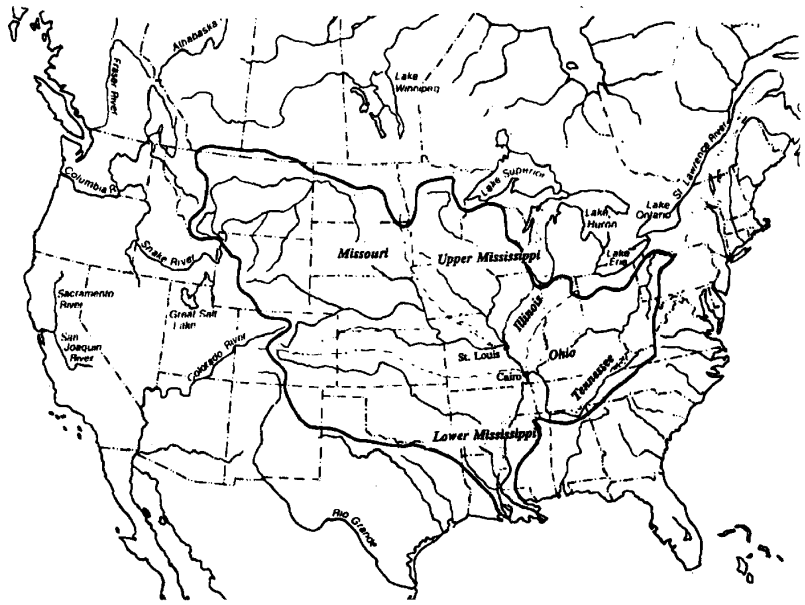


Figure 1 Mississippi River System and its main rivers.

Missouri. The areas most prone to large-scale simultaneous floods are the Ohio River and Lower Mississippi.

Human Setting

Today's human setting is a result of the basin's settlement patterns and the ensuing human development and use of the land and the river system. The original 13 American states claimed lands extending west to the Mississippi River well before any settlers had moved westward from the nation's east coast. As the eighteenth century ended, rapid settlement of the area began by westward movement into the Ohio River basin. The Louisiana Purchase of lands west of the Mississippi River from France for $15 million in 1803 doubled the size of the United States, surely ranking as one of the greatest bargains in history, and encompassed the entire Mississippi basin west of the river including the Missouri, Red, and Arkansas rivers.

This acquisition brought a rapid influx of settlers up the Mississippi River past New Orleans, and the formation of a major port at St. Louis. This invasion and settlement of the American center largely occurred during the first 75 years of the nineteenth century. The basin's estimated population in 1800 was 0.2 million and by 1890 it was 28 million. The Mississippi River and its tributaries were the "avenues" of settlement. They rapidly became the avenues of commerce to handle the flow of

goods in and products out. This led to the development of steamboats, and river transport became the first major commercial use of the river, forever establishing navigation as a major priority for the Mississippi River system (Interagency Committee, 1994). In the 1900s, Congress directed the U.S. Army Corps of Engineers to dredge a channel 6 ft deep from the mouth of the river to Minneapolis, and by 1950, a system of 29 locks and dams had been built along the Upper Mississippi from St. Louis to Minneapolis (Keating, 1971). In 1945 Congress authorized development of a 9-foot navigation channel for navigation on the Missouri River from St Louis to Sioux City, Iowa, with construction of six locks and dams (Interagency Committee, 1994).

Initial settlement involved farming in the fertile floodplains of the basin, followed by settlement of the uplands, which were largely prairies. Forests were cut to satisfy demands for wood of growing cities and to access new farmlands, an action that changed the landscape and made it more flood prone and erosion prone. Farming in the humid basins of the Mississippi and Ohio Rivers required extensive drainage works to eliminate the swamplike prairies, and this also changed the basin's water balance and further enhanced the movement of water and flooding. Farming in the more arid High Plains suffered from lack of water and ultimately led to widespread irrigation using river waters and groundwater. Congress passed the Reclamation Act of 1902 to aid irrigation in the West, and by 1990 more than 30,000 km^2 of the Missouri basin were being irrigated, further changing the region's water balance. As farming grew, towns developed along the rivers and the major port cities grew ever larger to handle the commerce of the basin.

By 1880 a settlement pattern involving intensive farming, transportation networks (by then roads and railroads as well as river transport), and cities with industry had emerged. The occurrence of floods within this now well-developed human setting brought chaos—people drowned, crops were washed away, and property destroyed, particularly along the floodplains of the great Mississippi. Action to address the floods of the Mississippi and its tributaries became a major issue for all levels, from local to federal governments.

2 EFFORTS TO CONTROL FLOODING: 1851 TO PRESENT

Levees and Warnings

Extensive floods in 1849 and 1850 (followed by an all-time record flood in 1858) led to action in Washington. Congress established the Delta Survey in 1851 to address the design and construction of works to control floods *and* to aid navigation on the Lower Mississippi River. The river's massive 1858 flood gave the survey team of army engineers a benchmark to work from. The Delta Survey team recommended a "levee-only" policy in 1861, a policy followed well into the twentieth century. The levee-only approach was to be done primarily to protect cities and communities along the river's main course. The U.S. Department of the Army was given the

primary responsibility for addressing the problem, and its engineers launched a program to control flooding.

In 1871 Congress directed the secretary of the Army to establish a network of river-stage gages along the Mississippi and Ohio Rivers. The Weather Service, established in 1870 as an arm of the Army's Signal Service, formed in 1873 a River and Flood Service, which began collecting the stream-level values and rainfall amounts (by telegraph) to make flood warnings, issued in Washington and transmitted by telegraph to district offices. These were issued for major flooding developing in the Lower Mississippi River basin, which was then the most flood-prone part of the entire Mississippi basin (Morrill, 1897). Flood prediction remained an ever-improving art based on empirical relationships of past rainfall conditions and river stages until the 1950s, when the science of rainfall forecasting had advanced. Such advances had led to improvements in flood forecasting on the Mississippi and its headwaters.

Most federal attention to flood control in the nineteenth century went to the construction of levees along the Lower Mississippi, then an easily flooded alluvial valley. In 1879 Congress established the Mississippi River Commission to survey the *entire* river system and to develop plans for navigation and flood control on all main river channels, reflecting growing federal responsibility for control of flooding. By this time, privately funded flood damage reduction measures, and mainly levees of varying types, were being built along parts of the Mississippi River. (Fig. 2).

As the nineteenth century ended, a system of major levees was being erected along the Lower Mississippi without any central planning or direction, and flooding continued there. Devastating floods in 1903 and 1912 led to ever more public

Figure 2 Refugees of the 1893 flood along the Lower Mississippi River. They survived by living on one of the many dirt levees built during the 1866–1890 era. Unfortunately, this scene has been repeated several times during the twentieth century.

pressure for government action against floods. Congress passed a Flood Control Act in 1917, calling for levee construction based on cost sharing with local districts, one of the first government–private sector partnerships.

Levees continued to be built in the basin, and in 1927 the Mississippi River Commission proudly announced that the levee system of the Lower Mississippi was ready to withstand the worst of floods. Two months later nature refuted their claim. An enormous spring flood developed on the Ohio River and spread into the Lower Mississippi River basin. It overwhelmed the massive levee system built during 1870 to 1926. The flood killed 246 persons and drove 600,000 from their homes as it spread over a 200-km wide swath extending from Cairo, Illinois, southward for 1500 km (Keating, 1971), covering large parts of Kentucky, Tennessee, Mississippi, Arkansas, and Louisiana.

Federal Preeminence

The shocking enormity of the 1927 flood should have revealed the fallacy of the levee-only policy, but the failure was blamed on God and the private district levees. This led Washington politicians to enact the Flood Control Act of 1928. It called for control of flooding on the *entire* Lower Mississippi River system. Major floods recurred in 1936 and 1937 (Smith, 1937), and these again led to more Congressional activity, actions that clearly established the federal government as being primarily *responsible for planning and accomplishing flood control across the nation.*

Fortunately, by the 1940s the human learning curve had led to a much better understanding of floods and how to manage floodplains. This included recognition that flood control works such as reservoirs were needed, and that these should serve multiple purposes including flood control, river navigation, water supply, and in later years, recreational needs (Wright, 1996). Hence, between 1936 and 1952, Congress spent $11 billion for flood control projects, primarily designed to store water. The Corps of Engineers built 76 reservoirs in the Upper Mississippi River basin and 49 on the Missouri River basin, and the Bureau of Reclamation built 22 flood control reservoirs in the Missouri basin. The New Deal government committed sizable funds to flood control in the Mississippi River basin and among its efforts was the establishment of the Tennessee Valley Authority in 1933.

Another policy shift in flood mitigation had evolved over time and eventually brought a development of a proper balance in the sharing for flood mitigation activities involving the states, the federal government (in charge), and local entities such as communities and flood control districts. A regional-scale approach to flood control also emerged under the leadership of U.S. Department of Agriculture's (USDA's) Soil Conservation Service in the 1940s. It had three regional components: (1) land treatment at the farm/local level (such as terracing), (2) upstream watersheds with flow retardation and channel stabilization, and (3) the standard downstream flood control measures (levees, pumps to remove water protected by levees, and major reservoirs). Efforts to improve flood mitigation were evolving.

Billions of dollars had been spent between 1851 and 1950 to structurally control flooding on the Mississippi River system, but it had not succeeded in abating flood

losses. The words of a well-known river sage, Mark Twain, issued 40 years before flood experts and Congress recognized their mistakes over 100 years, are relevant. In his *Life on the Mississippi* published in 1896, Twain wrote,

> Ten thousand river commissions, with the minds of the world at their back, cannot tame that lawless stream, cannot curb it or confine it, cannot say to it, "Go here," or "Go there," and make it obey; cannot save a shore which it has sentenced; cannot bar its path with an obstruction which it will not tear down, dance over, and laugh at.

Flood mitigation had long been a national goal, and by 1940 had become a federal responsibility. By then flood mitigation embraced a complex of various federal and state agencies, each with a different mission and with each often addressing varying constituents with conflicting views about how to manage the rivers and handle flooding.

These groups included farming interests, agribusiness, river transportation, hydroelectric power generation, irrigation, and recreation interests.

Changing Approaches for Handling the Flood Hazard

National recognition emerged in the 1950s that the structural approach to flood control was inadequate (White, 1958). This led to the development of a new thrust based on floodplain management through altering land use in floodplains and use of flood insurance, or "working with the river." The National Flood Insurance Program was established by Congress in 1968. Emerging environmental concerns in the 1960s led to the National Environmental Protection Act (NEPA) in 1968, bringing environmental quality into the objectives of water and floodplain management. This new "nonstructural" comprehensive approach to mitigating flood losses has evolved over the past 30 years but not yet solved the problems—flood losses continued to grow (National Science Foundation, 1980). Major floods on the Mississippi system occurred in 1965, 1973, and 1982–1983, and flash floods killed 236 in 1 hour at Rapid City, South Dakota, in 1972, and 139 in 2 hours in Colorado in 1976. As a result, many new projects for dealing with flash floods emerged.

Relief payments to flood victims, becoming an alternative to mitigation since 1970, became an ever-increasing way to address flood damages. Special relief payment legislation was issued by Congress after each major flood during the past 20 years. This mixed approach, relief and nonstructural, has essentially replaced the expensive flood control construction program of the 1851 to 1950 period. The enormous relief costs of the 1993 flood finally brought this budget-busting problem to the forefront, leading to badly needed changes in crop insurance and flood insurance programs (Changnon, 1996b).

Government policies relating to the struggle between humans and nature in the Mississippi River system have changed immensely over the past 150 years, but none have managed to solve the flood problem in the Mississippi River system.

3 IMPACTS FROM FLOODING

The failure of 150 years of national policies to solve the problems of flooding in the Mississippi River system was clearly illustrated by the impacts of three recent floods.

- A major summer (June and July) 1993 flood affected the Upper Mississippi and Lower Missouri Rivers (Changnon, 1996a), leading to $18 billion in losses and responses.
- In July 1996 a record-setting 43 cm rainstorm in 24 h occurred in Illinois with a large flash flood covering 15,000 km^2 in a matter of hours, engulfing a third of Chicago's suburban area in the headwaters of the Illinois River, a tributary of the Mississippi (Fig. 1).
- In 1997, a massive early spring flood occurred along the Ohio River, inundating many areas considered flood proof in five states.

Assessment of the impacts from recent flooding on the Mississippi River system drew heavily on events with these three recent, yet different types of floods. Major impacts were found in four broad sectors: (1) economic impacts, (2) environmental effects, (3) impacts to and responses by government at the local, state, and federal levels, and (4) social disruption.

Overview of Floods

The 1993 flood rated physically as the worst on the Upper Mississippi and Missouri Rivers, and the losses and costs of responding to the flood made it the most expensive flood on the Mississippi River system. Major losses included 52 dead, the highest since the floods of 1951, 56,000 homes damaged, 8.5 million farm acres either unplanted or unharvested, crop losses equating to $1.4 billion in corn and soybeans, and more than $1.9 billion losses to transportation systems including $920 million to the barge industry (20% of the year's revenue). The total losses amounted to an estimated $18 billion. Congress ultimately authorized $6.2 billion in aid. Insured crop losses totaled $1.6 billion, and the government paid out $301 million in flood insurance payments. State governments spent an estimated $1 billion in flood-related costs (Changnon, 1996a).

The large flash flood in July 1996 set record flow records on local streams and the Illinois River. It inundated parts of 18 large suburban communities of Chicago, causing enormous damage, disrupting transportation to Chicago, rural crop losses, and killing 5 people. Costs of the damages and responses to this flood exceed $0.6 billion (Changnon et al., 1997).

The late winter/early spring flood of 1997, due to heavy prolonged rains falling on frozen ground, did not set records. It was the most severe flood along the Ohio River since the floods of 1936, a 51-year period during which many thought that the area had become essentially flood proof. There were 24 killed with 83,000 homes damaged, and extensive damages to floodplain properties totaling $1.6 billion.

Economic Impacts

The economic impacts of the floods of 1993, 1996, and 1997 involved losses to individuals in and near flooded communities, to floodplain farmers, and to Midwest businesses and industries. Business losses affected regional sales, agricultural production, utilities, manufacturing, transportation, tourism, and recreation. However, the 1993 and 1997 floods also produced some winners in the agriculture, business, and transportation sectors.

Estimates of losses and costs of responding to the 1993 flood varied from $12 to $15.7 billion, and when the railroad and barge losses of $1.3 billion are added, the grand total is $18.1 billion, making it the nation's second worst weather disaster behind hurricane Andrew. The 1996 flash flood costs were $0.6 billion, and the estimates of costs of the 1997 flood are at $1.6 billion and growing.

The 1993 flood inundated vast amounts of valuable farmland representing about 4% of the Corn Belt's planted acreage. Lands lost to crop production due to the 1993 flooding included 2.5 million acres of corn and 1.97 million acres of soybeans. Agricultural losses amounting to $8.9 billion exceeded the losses of all other sectors. National corn production for 1993 was 31% less than for 1992, largely due to the flood. The effects of the 1997 flood on agriculture were less, since it occurred before planting had begun. Many nonflooded Midwestern farmers in 1993 came out "winners," as the flood caused grain prices to rise.

One of the greatest problems caused by all three floods was the curtailment of transportation. The 1993 flood became an absolute barrier to cross-river train and vehicular traffic, paralyzing transportation along 500 miles of the Mississippi River for 6 weeks. River-based barges were halted for nearly 2 months as about 1000 miles of navigable rivers were ultimately closed to navigation. With losses in excess of $1.9 billion, damages to the region's surface transportation systems were the worst in history. Barge movement along the Ohio River was halted for 5 weeks during the 1997 flood, a loss of $600 million. Revenues lost by navigation interests during the 2-month shutdown in 1993 amounted to $320 million.

The nation's major east–west railroads interconnect the top three rail hubs at Chicago, Kansas City, and St. Louis, but unfortunately cross through the badly flooded 1993 areas of Missouri, Iowa, and Illinois. Major washouts of track as well as bridge closings occurred in all floods and total damages for the railroads amounted to $240 million in 1993, and $38 million in 1996. Nearly 3000 long-distance trains had to be re-routed on longer, circuitous routes around the flood-affected areas. The damages and losses suffered by Midwest railroads ranked the 1993 flood as the worst natural disaster ever experienced by the railroad industry.

The public sector of surface transportation was heavily damaged in all three floods. Approaches to highway bridges were flooded and damaged. Road and highway closings during the floods were mainly concentrated along the major rivers, and in 1993, 3200 miles of roads were closed. The 1997 flood closed 75 highways in Ohio for at least 10 days. Severe erosion and washouts affected state and interstate highways, and hundreds of county roads, and damages amounted to $434 million in 1993, $78 million in 1996, and $185 million in 1997. Numerous floodplain busi-

nesses and industries were flooded or their operations were severely curtailed. Facilities were damaged, and production either stopped or slowed down greatly. Severe limitations on the transportation systems interrupted incoming and outgoing raw materials and manufactured products, producing loss of revenue and work stoppages. About 1900 businesses reported closings due to flooding in 1993, and as a result of the closures and commuting problems, 20,000 persons became unemployed.

Environmental Impacts

Many impacts on the environment were difficult to measure, some remain unmeasured, and many are tertiary and will take many more years before they are fully evident. The 1993 flood sizably altered the natural ecosystem of the Upper Mississippi and Missouri Rivers and their floodplains, changing many environmental conditions forever. The 1997 flooding also did enormous environmental damage in the floodplains in Kentucky, Ohio, Indiana, and Illinois. The 1996 flood did only minimal environmental damage with soil erosion and pollution of surface waters being the biggest impacts. Along with the flooding and related excessive erosion came further erosion and extensive silting to the floodplains and their wetlands. Although the two big floods (1993 and 1997) damaged some trees and plants, they generally provided a windfall for most plant and animal species, especially fish populations. Prolonged immersion of the nonfarmed portions of the floodplains had deleterious effects on certain trees. When levee breaks suddenly inundated vast areas, some wildlife already isolated by the flood drowned. Populations of certain insect pests were altered, at least on a 1-year time scale.

With more than 1000 levees failing in the Midwest in 1993, turbid and sediment-laden water moved out of the river into the newly exposed floodplains. Many backwater lakes along the Mississippi and Missouri Rivers had already lost between 70 and 100% of their capacities and lost substantially more volume after receiving sizable quantities of sediment during the 1993 flood. These excessive silts and sand in the floodplains smothered vegetation and compromised large areas of productive farmland.

Floodwaters with high flow rates resulted in much-above-normal amounts of eroded sediments and agricultural chemicals in the rivers, and these impacts from all three floods were sizable on water quality. The daily load of atrazine passing in the Mississippi River near Cairo, Illinois, in July 1993 was 12,000 lb per day, four times higher than during any previous year. A large percentage of the eroded herbicides and nitrates entering Midwestern rivers was carried into the Gulf and had a major impact on the ecosystem of the Gulf shore area. Another water quality problem in 1993, 1996, and 1997 along the rivers was large amounts of raw sewage, bacteria, viruses, and parasites carried by the floodwaters. Health officials were concerned that the organisms in the water could cause hepatitis, cholera, typhoid, or gastrointestinal illnesses; therefore, thousands of persons living and working along the river were inoculated to prevent disease outbreaks and, fortunately, waterborne diseases were minimal in all three floods.

Many parts of the ecosystems in and around the flooded rivers of 1993 and 1997 derived benefits from the floods. Large river–floodplain ecosystems in the river system have adapted to exploit seasonal flooding. A major problem in the 1993 flood related to Zebra mussels, which were inadvertently introduced in 1986 to the Great Lakes. The mussels entered the Illinois River from Lake Michigan via the canal system at Chicago in 1991 to 1992 and established themselves in the Upper Illinois River. These mussels released their larvae as the 1993 flood was occurring, and the floodwaters transported huge numbers of the larvae into the lower Illinois River and downstream into the Mississippi, moving laterally into many floodplain lakes and up many tributaries, and into industrial and municipal treatment plants. The Zebra mussel has prospered in its newly colonized habitats, adding greatly to the cost of water treatment and plant maintenance, jeopardizing the survival of native mollusks, and even altering river food webs by filtering detritus, suspended sediment, and the contaminants associated with these particles. This flood-induced spread of Zebra mussels was truly an "environmental disaster."

Impacts to Government

Government entities at all levels from local to federal levels experienced severe impacts due to the flooding. Many government activities fell within the broad definition of "responses," but many others were more "impacts." In 1993, 532 counties were identified as federal disaster areas; 11 counties in 1996; and in 1997, 79 counties were similarly declared. The federal government ultimately paid $6.2 billion for flood aid, insurance, and loans in 1993, and the total for 1997 is estimated at over $0.4 billion. Certain state agencies were heavily involved in flood and water monitoring, in emergency services, levee repair (National Guard units), water quality assessments, and in measuring the losses, representing a severe impact on state budgets, and the flooded states in 1993 spent an estimated $730 million on aid and rebuilding costs.

Many communities lost their water treatment plants for several weeks, making it extremely difficult and expensive to provide potable water, and many communities had severe or total losses of their sewage treatment plants. Mud-covered, flooded streets and city facilities required costly cleanup efforts, and the net result of the urban problems left many communities broke. Flood-fighting efforts at mid-sized river communities such as Quincy, Illinois, in 1993 cost $0.5 million, and $0.3 million in Aurora, Illinois, in 1996. Several flooded towns in 1993 considered relocation, and five of the badly flooded communities have subsequently relocated to higher ground.

Federal flood policies were impacted. The magnitude and damages of all three floods raised fundamental questions about the nation's floodplain management approach and the utility of the flood insurance program. In all three floods, the percent of those with floodplain insurance and damaged property was 10 percent or less. Excessive levee damages affected various governmental bodies. The levee system along the 1993 flooded rivers included 229 federal levees (39 damaged), 268

nonfederal levees (164 damaged), and 1079 private levees (879 damaged). The rebuilding of these levees has represented substantial costs to the federal government, to state governments, and to numerous local flood protection districts. Severely questioned were the benefits and effects of the development by the Corps of Engineers of the lock-and-dam system and the levee system. The Corps of Engineers calculated that the flood protection works (reservoirs and levees) on the Upper Mississippi had actually prevented an additional $4.9 billion in damages in 1993. Environmentalists countered, arguing that had the floodplains largely been left in their natural state, the two floods would have been of lesser magnitude and the damages due to unwise occupancy of the floodplains would have been negligible.

Social Disruption

The descriptions of the environmental effects, the sizable and pervasive economic impacts, and the complex maze of governmental actions due to the three recent floods all lead to the same obvious conclusion: There were considerable impacts on society in the flooded areas. Fatalities in the floods were a relatively small number considering the magnitude of the floods, reflecting improvements in flood prediction and warning. Fighting the flood was one of the major efforts of the 1993 and 1997 floods. The massive efforts involved thousands of persons residing in the threatened floodplains, volunteers, and hundreds of National Guard troops. Thousands of people were evacuated from their homes along the Mississippi in 1993, along tributaries of the Illinois River in 1996, and along the Ohio in 1997.

Anxiety among flood victims was high for long periods due to the initial fear of being flooded, the actual flooding and damages to personal property, and finally the exhaustive cleanup and restoration process. Loss of residence, or fear of its loss, was a primary cause of stress, along with loss of primary services, including protracted outages of power, water, and sewage treatment in communities and farms along the flooded rivers. Social disruption from the flood is most startling when viewed through the following numbers: of 94,000 persons evacuated from their residences in the summer of 1993, 45,000 were homeless at the end of November 1993, and 3000 were still homeless in June 1994. Furthermore, 61,000 Midwestern homes were seriously damaged, of which 60% were a total loss. The 1996 flash flood led to evacuations of 13,000, and 35,000 homes were flooded. The 1997 flood led to evacuations of 28,000 persons and flooding of 83,000 homes.

4 LESSONS

Assessment of the three recent floods led to the identification of major issues and common lessons about floods and their mitigation. All the issues and lessons learned for the 1993 flood have been defined in detail (Changnon, 1996b).

Floods Exceeding Past Experience and Design Extremes Continue to Occur

These extreme events caused unusual effects on riverine systems, extreme damage to "containment" structures, unexpected social and economic impacts, and assessments of the "cause" of the events came under scrutiny. Many system failures due to the floods were no one's fault—the design values were simply exceeded by conditions never or very infrequently experienced since river records have been kept.

Several scientific and technical actions are needed to improve understanding, mitigation, and response to extreme flooding. They include: (1) development of plans for data collection during and after floods, (2) development and installation of better instruments to measure floods and river flows, (3) development of hydrologic models for floods, and (4) timely collection of flood data before it disappears.

Major Unexpected Impacts Occurred

- *Unique Impacts to All Forms of Transportation* The nation's surface transportation systems, particularly the railroads and highway systems, experienced unusual and extensive damages from these three floods. The barge industry and shippers who depend on commercial navigation should seek improved river forecasting models and flood-monitoring systems. Approaches to many critical highway bridges need to be rebuilt to higher levels.
- *Structural Damage Exceeds Expectations* These floods with record rains and river levels offer lessons and information for engineers and structural experts about how to design structures more effectively to withstand flood extremes and to improve building codes. Current damage estimation techniques are inadequate. Data from the floods should be used to develop better guidelines for estimating flood damage.
- *River–Floodplain Ecosystems were Surprise Beneficiaries* Major floods, regardless of the human alterations in the floodplains, enhance river–floodplain ecosystems.
- *Human Actions Create Major Unexpected Environmental Problems* Human activities have hurt river ecosystems in many ways, and floods facilitate pest invasions and help create environmental disasters. The potential impacts of the nutrients and herbicides swept into the Gulf of Mexico in 1993 and 1997 need monitoring.
- *Unusual and Unplanned Adjustments and Responses Occurred* In extreme events, unexpected major impacts occurred and some existing governmental systems responded quickly and effectively. Ingenuity and resources are important ingredients in responding to extreme flooding.
- *Hopes for Restoring River Habitats in the Aftermath of the 1993 Flood Look Glum* The Corps of Engineers' annual budgets show sizable growth in funds for construction ($803 million in FY96, $1031 million in FY97, and to $1393

million in FY98), whereas funds for the Environmental Management Program of the Upper Mississippi declined from $19.5 million in FY96 to $12 million in FY98. Structural needs continue to overwhelm environmental concerns (Vanderpool, 1997).

Systems for Monitoring and Predicting Flood Conditions Were Inadequate or Failed

Existing systems for flood monitoring and flood forecasting are still inadequate. The National Weather Service needs better quantitative precipitation forecasts for periods 2 to 7 days ahead, and needs to more effectively integrate its new radar network to implement better flash flood warning techniques. The inadequacy of river monitoring equipment in the Mississippi River system calls for major improvements. Basin hydrologic models used for forecasting need revisions. There is a need for a layered geographical information system (GIS) for every river mile to allow better damage estimates as floods develop.

Flood Information Was Often Incomplete, Incorrect, or Not Timely

- *Loss Values* Data on flood conditions and losses were typically poor and generally inaccurate (often on the low side), and estimates remained highly inaccurate for considerable time after the floods. Means for obtaining more accurate near real-time data on conditions and losses should be developed to improve planning for in-flood adjustments and for relief and restoration activities.
- *Forecasts by Government Agencies* The operational hydrologic models used for flood predictions on all time scales need major improvements. Interactions between forecasters and hydrologists need improvement with a clarification of responsibilities. Methods used to estimate regional and national effects of large-scale wet and dry weather conditions on crop yields are inadequate and make sizable errors in growing season flood situations.
- *Confusion over Government Relief* Near real-time estimates of flood losses and predictions of the flood's size, both physically and economically, were underestimated in 1993 and 1997. They reflect the lack of real-time information about the magnitude of the flood and its impacts, plus poor outlooks about the growth of damages. The government should improve its means for acquiring information on impacts and work to remove or clarify overlapping responsibilities between agencies for handling relief aid for problem areas such as home reconstruction and levee rebuilding.
- *Public Understanding about Floods* There is widespread misunderstanding about floods and their frequency. A flood-related educational program would bring rewards in understanding forecasts, warnings, and description of terms used by scientists and engineers, plus clearer recognition of the risks related to living and farming in floodplains. Government officials need to realize how

affected citizens get disaster-related information (largely via TV) and utilize the broadcast media more effectively to disseminate information. The media has become the major source of information to victims and others interested in this form of natural hazard.

Many Approaches to Mitigate Flood Damages Failed But Some Succeeded

Government Flood Mitigation Policies Failed

Past structural and nonstructural approaches to flood mitigation have not worked, and major past efforts to improve U.S. flood policies have not succeeded. Only 10% of those flood damaged had flood insurance. The floods re-enforced the need to make improvements in floodplain use policies and the federal insurance programs. The considerable failure of the levee systems in 1993 and 1997 revealed that not all levees, particularly agricultural protection levees, can be built in a cost-effective manner to withstand floods. However, many past investments in flood control structures had utility.

Floods Produce Benefits

The major theme of these three floods was extensive losses. However, most weather events, including extremes such as droughts and floods, produce winners as well as losers. The 1993, 1996, and 1997 floods were not exceptions. Scientists and engineers benefited from new knowledge about floods, certain environmental problems are receiving needed attention, most aspects of the river–floodplain ecosystem benefited, inadequate federal policies gained public and political awareness and improved policies resulted, damaged often aged or inadequate facilities are being replaced by better facilities and equipment, and many farmers and businesses in nonflood areas benefited financially. The 1993 flood and its uniqueness produced major changes in flood policy including changes in the National Flood Insurance Program Act and the Federal Crop Insurance Program.

5 SUMMARY

Have the lessons taught by recent floods on the Mississippi River system been learned? The new (1994) law relating to flood insurance has not changed coverage purchased in areas flooded in 1996 and 1997. Obviously, floodplain residents continue to rely, as in the past, on "relief" as their "insurance" against floods and other hazards. Changes in the crop insurance laws in 1994 have led to increased purchases with less reliance on relief payments for flooding. Participation by at-risk populations will determine the extent to which future floods (certain to occur) will be damaging. Hopefully, the public will assume more responsibility for their actions.

Will changing government policies relating to reducing federal expenditures and focusing on more responsibility in the states and private sector help or hurt flood mitigation? In an era when cutting back government spending seems to be what the voter sees as prudent fiscal policy, are the cutbacks going to reduce investments in flood mitigation measures that may save inhabitants of the Mississippi River system substantial losses in the longer term? The resolution of this issue depends on public involvement and political wisdom and will. It remains to be seen whether government and the general populace will act on the lessons learned by the recent severe floods.

REFERENCES

Bhowmik, N., Physical effects: A changed landscape, in *The Great Flood of 1993*, Westview, Boulder, CO, 1996, pp. 101–131.

Changnon, S. A., W. C. Ackermann, G. F. White, and L. Ivens, *A Plan for Research on Floods and Their Mitigation in the United States*, Contract Report 302, Illinois State Water Survey, Champaign, IL, 1983.

Changnon, S. A., Losers and winners: A summary of the flood's impacts, in *The Great Flood of 1993*, Westview, Boulder, CO, 1996a, pp. 276–299.

Changnon, S. A., The lessons from the flood, in *The Great Flood of 1993*, Westview, Boulder, CO, 1996b, pp. 300–320.

Changnon, S. A., J. Angel, D. Changnon, F. A. Huff, P. Merzlock, P. Silberberg, and N. Westcott, *The Record Floods of July 1996 in Northeastern Illinois*, Miscellaneous Report 345, Illinois State Water Survey, Champaign, IL, 1997.

Faber, S., and C. Hunt, River management post-1993: The choice is ours, *Water Resour. Update*, *95*, 21–25, 1994.

Interagency Floodplain Management Review Committee, (IFMRC) *Sharing the Challenge: Floodplain Management into the 21st Century*, IFMRC, Washington, DC, 1994.

Keating, B., *The Mighty Mississippi*, National Geographic Society, Washington, DC, 1971.

Morrill, P., *Floods of the Mississippi River*, Bulletin E, Weather Bureau, Department of Agriculture, Washington, DC, 1897.

National Science Foundation (NSF), *A Report on Flood Hazard Mitigation*, NSF, Washington, DC, 1980.

National Weather Service, *The Great Flood of 1993*, National Disaster Survey Report, NOAA, Washington, DC, 1994.

Shabman, L., Responding to the 1993 flood: The restoration option, *Water Resourc. Update*, *95*, 26–30, 1994.

Smith, W. D., *The Flood of January 1937 in the Ohio and Lower Mississippi River Basins*, Illinois Department of Public Works and Buildings, Chicago, IL, 1937.

Tarbuck, E. J., and F. K. Lutgens, *The Earth*. C. E. Merrill, Columbus, OH, 1984.

Vanderpool, G., River restoration in jeopardy, *Mississippi Monitor*, *1*, 1–5, 1997.

White, G. F., *Changes in Urban Occupancy of Flood Plains in the United States*, Research Paper 57, Department of Geography, University of Chicago, Chicago, IL, 1958.

White, C. L., E. J. Foscue, and T. McKnight, *Regional Geography of Anglo-America*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

Wright, J. M., Effects of the flood on national policy: Some achievements, major challenges remain, in *The Great Flood of 1993*, Westview, Boulder, CO, 1996, pp 245–275.

CHAPTER 41

DROUGHT IN NORTHWEST AFRICA

WILL SWEARINGEN, ABDELLATIF BENCHERIFA

1 INTRODUCTION

The northwest African countries of Morocco, Algeria, and Tunisia frequently experience drought*. This region, commonly called the Maghreb, is situated between the Mediterranean and the Sahara Desert on the southern margins of midlatitude storm systems. As a result, both the timing and total amounts of rainfall are extremely irregular. Precipitation levels are generally insufficient for reliable or prosperous rain-fed agriculture in most of the region. Reduced rainfall is caused, among other factors, by the cold Canary Current off the region's western shores, which induces atmospheric stability and decreases the potential for rainfall. High-pressure ridges periodically develop offshore during the autumn–spring rainy season, barring access to moisture-bearing storms. If these high-pressure ridges persist for extended periods, drought results.

Drought is the leading natural hazard in the region and occurs frequently in all three countries. For example, during the twentieth century, Morocco has averaged 1 year of agricultural drought every 3 to 4 years. Unfortunately, there is no detectable periodicity. Each of these Maghreb countries experiences roughly the same frequency of drought. However, drought in one country is often not correlated with drought in the other two countries. For example, in 1988, Morocco had the largest cereal harvest in its entire history (a record since surpassed) while Tunisia suffered its worst harvest in over 40 years owing to drought.

*"Drought" as used here refers primarily to agricultural drought and is assessed through the use of cereal production statistics. This research was supported, in part, by grants from the Human Dimensions of Global Change initiative of the National Science Foundation and the Program in Science and Technology Cooperation of the U.S. Agency for International Development.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

Drought has major socioeconomic significance in northwest Africa because rain-fed (nonirrigated) cereal cultivation occupies a predominant place in the region's agriculture. Since pre-Roman times, this region has specialized in production of cereal crops, mainly wheat and barley, though maize and other cereals (including oats, sorghum, millet, rye, and rice) are also cultivated. Cereal crops account for approximately 85% of the region's cropland and are primarily produced by rain-fed methods. Wheat and barley are the mainstays of the national diets in the region and are consumed primarily as bread and cous cous.

Drought in this region sharply reduces both cereal acreage and yields, causing total production to plummet. This poses a food security threat, particularly if drought continues through a second year. Typically, during a drought year, food shortages develop, cereal imports rise dramatically, herds perish or are slaughtered for lack of forage, many farmers temporarily abandon their land and migrate to the cities, and soil erosion and desertification increase. Finally, the affected country's economy suffers a recession.

Good cereal harvests in northwest Africa require adequate rainfall during *both* the planting season (normally from October to December) and subsequent growing season (which extends until harvesting between April and June, depending on the region). Poor harvests or crop failure can result from rainfall shortages during either season. Given the potential for extreme interannual variability in precipitation levels, 400 mm annual average precipitation is normally considered the threshold for viable rain-fed cereal production in northwest Africa (Bencherifa, 1988b). However, the temporal distribution of rainfall is just as critical as the total amount. For example, if the entire winter precipitation falls during a few intense cloudbursts, most will disappear as runoff and be unavailable for crop use. Thus, regardless of the total amount of rain, drought conditions will probably develop.

2 INCREASING VULNERABILITY TO DROUGHT

Since the earliest historical times, drought has been a major hazard in northwest Africa. Historical surveys of drought and other natural calamities have determined that there were 49 major drought-related famines in Morocco during the period from the late ninth century to the early 1900s (Bois, 1957) and at least 26 such episodes in Tunisia from around AD100 to the late 1800s (Bois, 1944).

While the drought hazard has perhaps always existed in northwest Africa, this hazard has been increasing during the present century (Swearingen, 1992, 1994, 1996a). It has been increasing primarily due to two key processes: (1) expansion of cereal cultivation to drought-prone rangeland and (2) reduction of fallow. During the colonial period, these processes were fostered by large-scale land expropriation, by the dislodging of peasants to marginal lands, by a cereal policy offering high crop prices and other incentives, by agricultural mechanization, which facilitated the mining of marginal areas during periods of higher-than-normal rainfall, and by population pressure associated with rapid population growth. Other significant factors during this period include the gradual loss of peasant ability to stockpile

grain as insurance against drought and the progressive substitution of wheat for drought-resistant barley. Since independence, populations in northwest Africa have continued to multiply at rapid rates. High population growth rates, along with neglect of cereal production, gradually precipitated a food security crisis by the early 1980s (Swearingen, 1987b, 1996b). To counter this crisis, all three countries have been making concerted efforts to boost their cereal production. Unfortunately, the policies adopted are further promoting cultivation of drought-prone rangeland and reduction of fallow.

The Colonial Period

Algeria became a French colony in 1830, Tunisia a French protectorate in 1881, and Morocco a French and Spanish protectorate in 1912. The colonial period lasted until 1956 in Tunisia and Morocco, and 1962 in Algeria. In all three countries, colonization introduced major changes. The net effect of these changes was a gradual increase in the drought hazard.

Prior to the colonial period, agriculture in Northwest Africa consisted of an extensive system of dry-land cereal cultivation and animal husbandry, with irrigated orchards and gardens surrounding most urban centers and many villages and pastoral nomadism practiced in the desert regions (Bencherifa, 1986, 1988a; Swearingen, 1987a). Most land was communally owned. Rain-fed landholdings were concentrated in higher rainfall areas with above 400 mm rainfall per year. Landholdings were usually dispersed to provide for equity and to help counter the risk of crop failure. Each peasant farmed several dispersed plots. Surplus grain from bountiful harvests was stockpiled to cover crop failures during drought years. Some stockpiling also occurred at the national level during precolonial times. For example, many of the Alawi sultans in Morocco maintained large granaries as a hedge against drought and famine (Meyers, 1981). Fallowing (periodically letting cropland lie idle instead of cultivating it) was widely practiced. Fallowing both replenished soil moisture and helped to restore soil fertility. Low population pressure gave the arable expanses a relatively underutilized appearance. In addition, lower-rainfall areas were used only for seasonal stockraising.

French colonial planners viewed northwest Africa as ideal for large-scale French settlement. In all three countries, colonization dislodged peasants from much of the best land. Europeans acquired roughly 30% of Algeria's arable land (or 2.7 million hectares), nearly 20% of Tunisia's land (or 800,000 hectares), and 15% of Morocco's land (or 1 million hectares).

Exacerbating the effect of European colonization was land concentration by native large landowners. During the colonial period, indigenous landowners allied with the French were able to amass sizable landholdings in all three countries. In Algeria, some 25,000 native Algerians acquired a total of nearly 2.8 million hectares, somewhat over 30% of the country's arable land (Pfeifer, 1985). In Morocco, 7500 Moroccan landowners acquired 1.6 million hectares, or 24% of the arable total (Swearingen, 1987a). And in Tunisia, 7200 Tunisians acquired 630,000 hectares—15% (Sethom, 1985).

Land concentration during the colonial period had two important consequences: First, as land was expropriated, peasants became concentrated on a diminished amount of land. This reduced peasants' ability to let part of their land lie fallow (Bencherifa and Johnson, 1990, 1991). Reduction of fallow significantly increased the potential for drought. The primary purpose of fallowing in semiarid regions like northwest Africa is to allow soil moisture to accumulate (WMO, 1975). Approximately 20 to 25% of the precipitation falling during the fallow rainy season (roughly October to April) is retained in the soil. Thus, fallowing substantially boosts the available water supply for subsequent crop use. In low-rainfall areas, this soil moisture component is often the critical difference between a successful harvest and drought. With the reduction of fallowing, this buffer was lost, and vulnerability to drought increased. In addition, excessive land-use pressure caused soil fertility to decline. The resulting impoverishment of their land made it increasingly difficult for peasants to stockpile grain as a hedge against drought.

Second, large masses of peasants were dislodged to marginal land that was not sufficiently attractive for colonization. The marginal areas were commonly characterized by poor soils, unfavorable slope, or deficient rainfall. Previously, most of this land had been used only for livestock grazing. Once under plow, it became prone to soil erosion and desertification. Unfortunately, it also became more vulnerable to drought (Bencherifa, 1996).

While land concentration was taking place during the colonial period, other significant changes were occurring. Health measures introduced by the French caused native death rates to plunge. Northwest Africa's population expanded dramatically, with roughly a fivefold increase during colonial times. This population explosion (combined with the expropriation of between a third and a half of the arable land by Europeans and indigenous large landowners) intensified pressure on remaining agricultural resources. Fallow was further reduced, peasant landholdings became increasingly fragmented, soil fertility continued to decline, and more marginal land was put under cultivation.

Colonial agricultural policy, per se, also played a major role in deepening northwest Africa's vulnerability to drought. Between roughly 1915 and 1928, colonial authorities in all three countries had a mandate from the *métropole* to substantially boost cereal production for France. The architects of this mandate were convinced that France's *Afrique du Nord* had been a bountiful breadbasket for Rome during classical times, and that France could restore this land to its former productivity (Swearingen, 1987a). Various subsidies and bonuses were offered to encourage cereal cultivation, especially cultivation by mechanized means. High market prices were also offered, particularly for wheat. Agricultural mechanization and high crop prices enabled marginal areas to be profitably cultivated during higher-than-normal rainfall periods. Although Europeans and native large landowners were the primary beneficiaries of the subsidies and bonuses, lucrative crop prices also encouraged peasant farmers to significantly expand their cereal acreage. Cereal acreage in the region increased dramatically.

Contributing to northwest Africa's vulnerability to drought was the fact that the colonial policy favored wheat production over barley. Previously, barley had been the

predominant native cereal. However, wheat now became predominant, and consumer tastes changed to prefer this cereal. With the varieties at the time, the critical rainfall limits for barley were some 30% less than those for wheat. In addition, barley ripens and can be harvested significantly earlier than wheat; therefore, it is less vulnerable to the untimely onset of summer drought conditions. In short, by substituting wheat for barley, the colonial wheat policy increased the potential for significant drought impacts.

Since Independence

Since independence, each of the northwest African countries has pursued a different development strategy. Algeria, emerging from a traumatic colonial experience and devastating war of independence in 1962, has attempted to achieve economic independence through a comprehensive program of industrialization. Morocco, since independence in 1956, has emphasized export agriculture, investing heavily in irrigated production of citrus and market vegetables. Tunisia has adopted the most balanced development strategy since its independence in 1956: it has invested in export agriculture and has also encouraged export-led industrialization by multinational firms.

All three countries recovered ownership of colonial landholdings and have engaged in limited land reform. However, much of the former colonial land passed into the hands of more prosperous native landowners. Furthermore, most of the large landholdings acquired by native landowners during the colonial period were never subject to land reform.

For at least two decades following independence, the northwest African countries seriously neglected domestic food production. By the early 1980s, all three countries were experiencing a food security crisis. Key symptoms of this crisis were declining per capita cereal production, alarming, ever-growing levels of cereal imports; heavy foreign indebtedness related to these imports, and massive food subsidy programs.

By the early 1980s, Algeria was importing approximately two-thirds of its cereal supply, Tunisia was importing nearly half, and Morocco was importing over a third (FAO, various years). In each country, a significant percentage of the population was having difficulty meeting daily food needs. The political implications of this crisis became clear by 1981, when Morocco experienced a bloody food-related riot. Similar food-related riots erupted in Morocco and Tunisia in 1984 and in Algeria in 1988.

Since the early to mid-1980s, all three countries have been undertaking major agricultural reforms (Swearingen, 1996b). The overriding objective is to increase dry-land cereal production. Reforms include privatization of the state agricultural sectors to improve efficiency and promotion of modern seed varieties and fertilizers. In terms of drought enhancement, however, the most significant reforms involve changes in crop prices, especially in Morocco, promotion of agricultural mechanization, and a "new lands" policy in Algeria.

Since independence, northwest African governments have maintained tight control over producer prices of basic food crops. Prices for these crops, cereals in particular, were held artificially low until the 1980s. Indeed, for much of this period,

crop prices were only about a fourth of what they would have been without government intervention (Cleaver, 1982). Government rationale was that low crop prices would enable them to provide cheap food to their urban populations, helping to keep wages low and thereby assisting industrialization and other urban development initiatives. An ulterior motive behind the cheap food strategy was to help prevent social unrest among the growing ranks of the urban poor. Unfortunately, low crop prices acted as a major disincentive to farmers, creating a vicious spiral of declining production.

Beginning in the late-1970s, fixed producer prices for cereals and other basic food crops were gradually raised. In Algeria and Tunisia, these prices approached world market levels by the mid-1980s, helping to stimulate cereal production and extension of cultivation to previously uncultivated rangeland areas. However, in Morocco, changes in pricing policy were far more dramatic. Here, the government boosted producer prices of barley and wheat to nearly *twice* world market levels. The stimulus effect has been remarkable and has led to a major expansion. Average annual cereal acreage during the 1980 to 1984 period was slightly over 4.4 million hectares. However, during the 1985 to 1989 period, it expanded to 5.2 million hectares—an increase of over 15% (FAO, various years). This increase has come primarily through the extension of cereal cultivation to marginal rangeland and reduction of fallow. Both of these processes are increasing vulnerability to drought.

Government efforts in all three northwest African countries to promote mechanization have facilitated the extension of cereal cultivation to drought-prone rangeland. The tractor and disc plow have colonized large stretches of rangeland in all three countries. In southern Tunisia, for example, roads created by oil exploration crews have enabled mechanized farmers to penetrate regions that previously were accessible only to pastoral nomads. Similar penetration of previously remote grazing lands has also occurred in the other two countries. Some of these new lands normally receive as little as 200 mm of annual rainfall. Their poor soils can sustain cultivation for a few years, as long as higher-than-normal rainfall prevails. However, the return of normal (reduced) rainfall forces their abandonment. Desertification quickly advances in the abandoned areas.

In Algeria, cultivation of marginal lands has actually become official policy. In 1983, Algeria's government passed legislation that established an ambitious homesteading program. The overriding purpose of this program is to encourage Algerian citizens to maximize the agricultural potential of the country through development of public domain land that has not previously been cultivated.

The government views the program as a way to expand the agricultural resource base, increase the food supply, combat peasant exodus to the cities, and counterbalance excessive urban development along the country's northern coast. The goal is to put approximately 800,000 hectares of new land into production. About half of this land will be in the Saharan zone and involves small (less than 3-hectare) irrigated plots. However, the other half, some 400,000 hectares, involves larger dry-land allotments in the country's high plateau region. Virtually all new "cropland" in this region is low-rainfall steppeland suitable only for stockraising. The

homesteading program, then, will significantly increase the proportion of Algeria's cropland in drought-prone areas.

The homesteading program, however, is only part of Algeria's current new lands scheme. In 1984, the Algerian government initiated a comprehensive agricultural plan that includes the goal of putting *2 million* hectares of new land into production. Two-fifths of this new land is to come from the homesteading program. The other three-fifths, or 1.2 million hectares, will come from reduction of fallow in the traditional crop rotation system. This major reduction of fallow, for reasons previously discussed, will substantially increase the risk of drought in Algeria.

3 FIELD RESEARCH TO ASSESS LINKAGES BETWEEN HUMAN ACTIVITIES AND DROUGHT

To help assess the linkages between human activities and drought in northwest Africa as well as the region's increasing vulnerability to drought, the authors organized an extensive field research project in Morocco during the early 1990s. A project team led by one of the authors (Bencherifa) intensively interviewed farmers about farming practices and drought-coping strategies. These interviews were conducted in three different regions of Morocco: the Chaouia (a subhumid region), the northeast, usually referred to as Maroc oriental (a semiarid region), and the Chichaoua (an arid region).

The research team surveyed a total of 335 households or production units. The survey consisted of a series of six questionnaire interviews, which were administered orally to the same household units over a period of nearly 2 years, from 1992 to 1994. The interview protocol was intended to capture information about dynamic responses of producers to specific climate conditions, including both drought and abundant rainfall. By coincidence, the survey covered periods of highly variable weather, including drought and higher-than-normal rainfall years.

In all three study areas, it quickly became clear that rainfall variability is well accounted for in production strategies. The historical backgrounds of the communities in these study areas provide, in effect, a reservoir of memories that allows drought to be regarded as a normal rather than exceptional event. Farmers expect periodic drought and plan for it in their production strategies.

However, the survey also revealed that traditional drought-coping strategies have been losing their effectiveness. Namely, the strategies have been weakened by increasing population pressure and more intensive use of agricultural land. In addition, increasing inequalities between producers (a result of the unequal penetration of market-oriented farming practices) have increased the vulnerability of the poorest farming households to the impacts of drought.

From this extensive field survey, the research team was able to make several generalizations, which can be extrapolated to northwest Africa as a whole:

1. *Vulnerability to Drought Is Related to a Variety of Agronomic Factors* Agronomic factors that help determine vulnerability to drought include the

following: (a) The actual time of plowing and planting: These operations need to be keyed to the timing of the first fall rains in the October to November period. Farmers need to make basic decisions about when to plant, which entails risks. Farmers who achieve optimal timing in planting are less likely to be impacted by drought than farmers who plant either too early or too late. (b) The specific crops grown: Barley is the most drought resistant cereal crop—thus its domination in arid and semiarid conditions. Hard and soft wheat are more sensitive to shortfalls of precipitation. (c) The preceding land use: Fallow helps to mitigate the effects of drought because of the accumulation of soil moisture in fallowed fields. (d) The amount and type of labor inputs: Labor inputs allocated for preparation of soils (most importantly, animal traction versus modern machinery) have a major influence on vulnerability to drought. (e) The type of soil: Heavy soils are excellent agronomically when rainfall is above average. However, light soils have advantages during years of below-average rainfall.

2. *Fallow Has a Critical Role within the Agropastoral System* Fallowing is universally recognized by farmers to be a major determinant in increasing crop output. During the fallow year (as previously noted), fields not only accumulate soil moisture but also nitrogen, leading to increased yields when these fields are again cultivated. In addition, fallow is an essential part of livestock production. This is because farmers obtain fodder both from stubble remaining from the previous year's cultivation as well as from weeds that grow on fallowed fields. When fallow disappears from the agropastoral system due to demographic pressure, farmers become more vulnerable to drought. In all three regions, fallow has been steadily decreasing, as is generally true throughout northwest Africa.
3. *Livestock Raising Is a Basic Drought-Coping Strategy* Livestock play a key role in farmer survival strategies. Despite environmental and demographic differences, most of northwest Africa is characterized by a combination of animal herding and cultivation. However, agropastoral systems differ considerably in herd composition. Herds typically vary from cattle, to combined cattle and sheep, to sheep and goats. Differences are due both to the availability of arable land versus rangeland (in part, the result of different levels of population pressure) and the quality of the rangeland. Because of its comparatively high income-generating potential, livestock production is regarded as the key way to maximize farm incomes during years of adequate or abundant rainfall. However, this integration of livestock production and cultivation increasingly is becoming dysfunctional in the case of multiyear droughts. This is because it relies heavily on the use of by-products from cultivation for animal fodder (hay, straw, stubble, and weeds from fields in fallow). Available fodder has largely disappeared from most farms following a single year of drought. Because of population pressure and the conversion of higher-quality rangeland to cultivation, northwest Africa's agropastoral system has become increasingly vulnerable to drought.

4. *Farmers Employ a Traditional Suite of Other Strategies to Cope with Drought* These include the following: (a) Grain and animal fodder are stockpiled using a variety of traditional storage systems, including conical stacks of hay and underground grain storage pits. Stockpiling grain and fodder is an effective way to buffer drought's impacts, particularly if it does not continue for more than a single year. (b) Farmers reduce their herd size to a level that can be sustained through the drought. However, even in severe drought conditions, they attempt to maintain a small herd of breeding stock. This core herd allows a new start once rainy conditions return. (c) Farmers often adopt a relatively mobile, pastoral-nomadic stockraising pattern to seek grazing resources elsewhere if they run out of fodder on their own farms. (d) Farmers take advantage of rainfall whenever it occurs. If the normal cereal crops cannot be planted in fall or early winter because of drought, and if late-season rain occurs, they plant late crops such as chickpeas or lentils.

   During the early 1990s in Morocco, farmers progressively adopted *new drought-coping strategies*, which can be found generally in northwest Africa. These included the following: (a) Farmers almost universally adopted mechanization wherever possible. When farm income did not allow them to purchase their own farm machinery, they hired plowing services. Mechanization of plowing, in particular, has become a general drought-coping strategy because it allows both for rapid planting following the first rains as well as for quick response to rainfall. For example, in case of late spring rain after previously planted crops have failed, modern farm machinery allows for rapid replanting. Mechanization also dramatically increases farm output during favorable rainfall years by allowing farmers to maximize the cultivated area. This increased production can be stockpiled as a hedge against future drought. However, mechanization also has increased the negative impacts of drought. In short, it has helped agricultural production become a "high risks, high rewards" game. (b) Farmers have adopted fertilizer use as a way to increase production during good years to stockpile grain and fodder in preparation for future drought. (c) Farmers have adopted intensive livestock raising in stables as a way to increase farm income. (d) Wherever possible, farmers have attempted to develop irrigation through digging of new wells, use of motor pumping, and use of diversion devices to concentrate runoff to their plots. (e) Farm families increasingly rely on off-farm income to supplement their farming resources. This strategy included temporary migration to urban areas by one or more family members during drought years. In the severe 3-year drought in the Chichaoua in Morocco during the early 1990s, around 80% of a typical family's resources came from outside the farm.

5. *Socioeconomic Impacts of Drought Are Related to Its Duration* The duration of drought is a major determinant of its socioeconomic impacts. One year of drought following a normal rainfall year has far fewer negative impacts at the household level than is commonly assumed. This is because of the effectiveness of traditional drought-coping strategies, including stockpiling of grain

and fodder during higher-than-normal rainfall years. The general calculation among the farmers surveyed is that "good years cover the bad years." Only when drought lasts more than a single year do its effects generally become critical at the household level. When drought lasts more than a single year, stockpiles of grain and fodder become exhausted, throwing household economies into crisis and threatening starvation for both livestock and people.

4 CONCLUSION

A highly vulnerable agricultural system has emerged in northwest Africa owing to historical, demographic, economic, technological, and policy-related factors. In all three countries, increasing population pressure and higher market demand have exerted pressure on local natural resources, resulting in the extension of cultivation to lower-rainfall areas and the reduction of fallowing. Data collected during extensive field research in Morocco suggest that most northwest African farmers can successfully endure only a single year of meteorological drought without significant hardship. If drought continues through a second year or longer, socioeconomic impacts become critical, even devastating.

REFERENCES

AID. Morocco: Country development strategy statement (FYs 1987–1991). Annex C: The agricultural sector in Morocco: A description, unpublished report, United States Agency for International Development, Washington, DC, February 1986.

Bencherifa, A., Agropastoral systems in Morocco. Cultural ecology of tradition and change, Ph.D thesis, Clark University, Worcester, MA, 1986.

Bencherifa, A., Agropastorale Organisations for men im Atlantischen Marokko, *Die Erde, 119*, 1–13, 1988a.

Bencherifa, A., Le Monde rural marocain, in T. Agoumy and A. Bencherifa (Eds.), *La Grande Encyclopédie du Maroc: Géographie Humaine*, GEP, Cremona, Italy, 1988b

Bencherifa, A., Is sedentarization of pastoral nomads causing desertification? The case of the Beni Guil of Eastern Morocco, in W. Swearingen and A. Bencherifa (Eds.), *The North African Environment at Risk*, Westview Boulder, CO, 1996, pp. 117–130.

Bencherifa, A., and D. Johnson, Adaptation and intensification issues in pastoral systems: Observations in Morocco, in J. G. Galaty and D. Johnson (Eds.), *The World of Pastoralism*, Guilford, New York, 1990, pp 394–416.

Bencherifa, A., and D. Johnson, Resource management changes in the Middle Atlas Mountains: From extensive pastoralism to intensive cash production, *Mountain Res. Devel. 3*, 183–194, 1991.

Bois, C., Années de disette, années d'abondance: Sécheresses et pluies en Tunisie de 648 à 1881, *Rev. lEtude Calamite's, 21*, 3–26, 1944.

Bois, C., Années de disette, années d'abondance: Sécheresses et pluies au Maroc, *Rev. lEtude Calamités, 2635* , 33–71, 1957.

Cleaver, K., *The Agricultural Development Experience of Algeria, Morocco and Tunisia: A Comparison of Strategies for Growth*, Staff working paper 552, World Bank, Washington, DC, 1982.

Food and Agriculture Organization (FAO), *Production Yearbook*, FAO, Rome, various years.

Meyers, A.R., Famine relief and imperial policy in early modern Morocco: The political functions of public health, *Am. J. Public Health, 71*, 1266–1273, 1981.

Morocco, *Consommation et Dépenses des Ménages 198485. Premiers Re´sultats*, Vol. 1: *Rapport de Synthése*, Direction de la Statistique, Rabat, 1988.

Pfeifer, K., *Agrarian Reform under State Capitalism in Algeria*, Westview, Boulder, CO, 1985.

Sethom, H., L'Action des pouvoirs publics sur les paysages et l'économie rurale dans la Tunisie indépendante, in P. R. Baduel et al. (eds), *Etats, Terrotoires et Terroires auMaghreb*, Centre National de la Recherche Scientifique, Paris, France, 1985, pp. 98–113.

Swearingen, W., *Moroccan Mirages: Agrarian Dreams and Deceptions, 1912-1986* , Princeton University Press, Princeton, 1987a.

Swearingen, W., Morocco's agricultural crisis in I. W. Zartman (Ed.), *The Political Economy of Morocco*, Praeger, New York, 1987b, pp. 159–172.

Swearingen, W., Drought hazard in Morocco, *Geogr. Rev. 82*, 401–412, 1992.

Swearingen, W., Northwest Africa, in M. H. Glantz (Ed.), *Drought Follows the Plow*, Cambridge University Press, Cambridge, 1994, pp. 117–133.

Swearingen, W., Is drought increasing in Northwest Africa? A historical analysis, in W. Swearingen and A. Bencherifa (Eds.), *The North African Environment at Risk*, Westview, Boulder, CO, 1996a, pp. 17–34.

Swearingen, W., Agricultural reform in Northwest Africa: Economic necessity and environmental dilemmas, in D. Vandewalle (Ed.), *North Africa: Development and Reform in a Changing Global Economy*, St Martins', New York, 1996.

World Meteorological Organization (WMO), *Drought and Agriculture*, Technical Note No. 138, WMO, Geneva.

World Bank, *Social Indicators of Development*, Johns Hopkins University Press, Baltimore, 1989.

CHAPTER 42

HURRICANE AS AN EXTREME METEOROLOGICAL EVENT

ROGER A. PIELKE, JR. AND ROGER A. PIELKE, SR.

1 INTRODUCTION: UNDERSTANDING SOCIETAL RESPONSES TO EXTREME WEATHER EVENTS

In the 1970s, many decision makers became increasingly interested in climate because of numerous weather-related impacts around the world. Events that helped to stimulate this interest included the failed Peruvian anchovy harvest in 1972 and 1973, the 1968 to 1973 drought in the African Sahel, a severe winter freeze in 1972 in the Soviet Union, and in 1974 floods, drought, and early frost in the U.S. Midwest. In 1977, winter in the eastern United States was the coldest ever recorded and summer was one of the three hottest in a century. As a consequence of these extreme events and their impacts, decision makers began paying more attention to the relation of weather and climate to human affairs.

Understanding societal responses to weather and climate requires an understanding of the terms *weather* and *climate*. The 1979 World Climate Conference adopted the following definitions of weather and climate:

> *Weather* is associated with the complete state of the atmosphere at a particular instant in time, and with the evolution of this state through the generation, growth and decay of individual disturbances.

> *Climate* is the synthesis of weather events over the whole of a period statistically long enough to establish its statistical ensemble properties (mean value, variances, probabilities of extreme events, etc.) and is largely independent of any instantaneous state.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

Climate refers to more than "average weather" (Gibbs, 1987). Climate is, in statistical terminology, the distribution of weather events and their component properties (e.g., rainfall) over some period of time, typically a few months to thousands of years. In general, climate statistics are based on actual (e.g., weather station) or proxy (e.g., ice core) records of weather observations. Such a record of weather events can be used to create a frequency distribution that will have a central tendency, which can be expressed as an average, but it will also have a variance (i.e., spread around an average). Often, variability is more important to decision makers than the average state (Katz and Brown, 1992).

How society thinks about "extreme" weather is, of course, related to what is defined as "normal" weather. What, then, is a "normal" weather event? There are different ways to define normal weather. Of course, it is possible to argue that on planet Earth all weather events are in some sense normal; however, such a definition has little practical utility for decision makers. One way to refine the concept is to define normal weather events as those events that occur within a certain range within a distribution, such as, for instance, all events that fall within one standard deviation of the mean. In practice, historical records of various lengths and reliabilities have been collected around the world for temperature, precipitation, storm events, and others. When data is available, such a statistical definition lends itself to equating normal weather with "expected" weather, where expectations are set according to the amount of the distribution defined as normal. For example, about 68% of all events fall within one standard deviation of the mean of a bell-shaped distribution.

A change in the statistical distribution of a weather variable—such as that associated with a change in climate—is troubling because decision makers may no longer expect that the future will resemble the past. For the insurance industry, as well as other decision makers who rely on actuarial information, such a possibility of a changing climate is particularly troubling. A climate change is thus a variation or change in the shape or location (e.g., mean) of a distribution of discrete events (Katz, 1993).

"Extreme" weather events can simply be defined as those not normal, however normal is chosen to be defined. For instance, if normal weather events are those that occur within 2 standard deviations of the mean, then about 5% of all events will be classified as extreme.

While it is possible to classify hurricanes as either "normal" or "extreme" in this manner, the simple fact is that for most communities any landfalling hurricane would qualify as an extreme event because of their rarity at particular locations along the coast.

From the standpoint of those human activities sensitive to hurricane impacts, it is often the case that decisions are made and decision processes established based on some set of expectations about what future weather or climate will be like. Building codes, land-use regulations, insurance rates, disaster contingency funds are each an example of decisions that are dependent upon an expectation of the frequency and magnitude of future normal and extreme events.

In short, decision makers typically establish policies based upon an expectation of normal weather. Yet for most coastal communities normal weather has historically

(or at least over the time of a human memory) meant no hurricanes! Consequently, people are often surprised when a hurricane does strike and then overwhelms response capabilities. Because decision makers do not always consider the possibility of extreme weather, when such events occur, they often reveal society's vulnerabilities and sometimes lead to human disaster. A fundamental challenge facing society is to incorporate information about weather and climate risks into decision making in order to take advantage of normal weather and to prepare for the extreme. The degree to which society exploits normal weather and reduces its vulnerabilities to extreme weather is a function of how society organizes itself in the face of what is known about various typical and extreme weather events. The challenge is made more difficult by variability at all measurable time scales in the underlying climate, and hence in the frequency, magnitude, and location of various weather events. And, of course, decisions that have a weather or climate component also are laden with all of the political, practical, and social factors that influence policy.

2 HURRICANES DEFINED

One of the most powerful natural phenomena on the face of Earth, the hurricane is a member of a broader class of phenomena called cyclones.* The term *cyclone* refers to any weather system that circulates in a counterclockwise direction in the Northern Hemisphere and in a clockwise direction in the Southern Hemisphere. "Tropical cyclones" typically form over ocean waters of the tropics. The tropics are the area on Earth's surface between the Tropic of Capricorn and the Tropic of Cancer, 23° 27′ south and north of the equator, respectively. Extratropical cyclones, for comparison, form as a result of the temperature contrast between the colder air at higher latitudes and warmer air closer to the equator. Extratropical storms form over both the ocean and land.

Tropical cyclones have been given different names depending on their region of origin. In the western north Pacific, they are called typhoons, while in the Bay of Bengal they are referred to as severe cyclonic storms of hurricane intensity. In the Atlantic, Gulf of Mexico, Caribbean, and Pacific north of the equator and east of the international dateline they are hurricanes. Evidence of tropical cyclones has been documented in a variety of other geographic locations including Europe and North Africa at earlier geologic times (Ager, 1993). Figure 1 shows the tracks of all tropical cyclones with winds greater than 39 mph for the 10-year period 1979 to 1988.

The meteorological community uses a number of terms to classify the various stages in the life cycle of tropical cyclones. The following are definitions of tropical cyclones used in the Atlantic Ocean basin (Pielke and Pielke, 1997):

*This chapter considers hurricanes as an extreme meteorological event. It first discusses the physical aspects of hurricanes, including their development and impacts on ocean and land. It then overviews societal impacts.

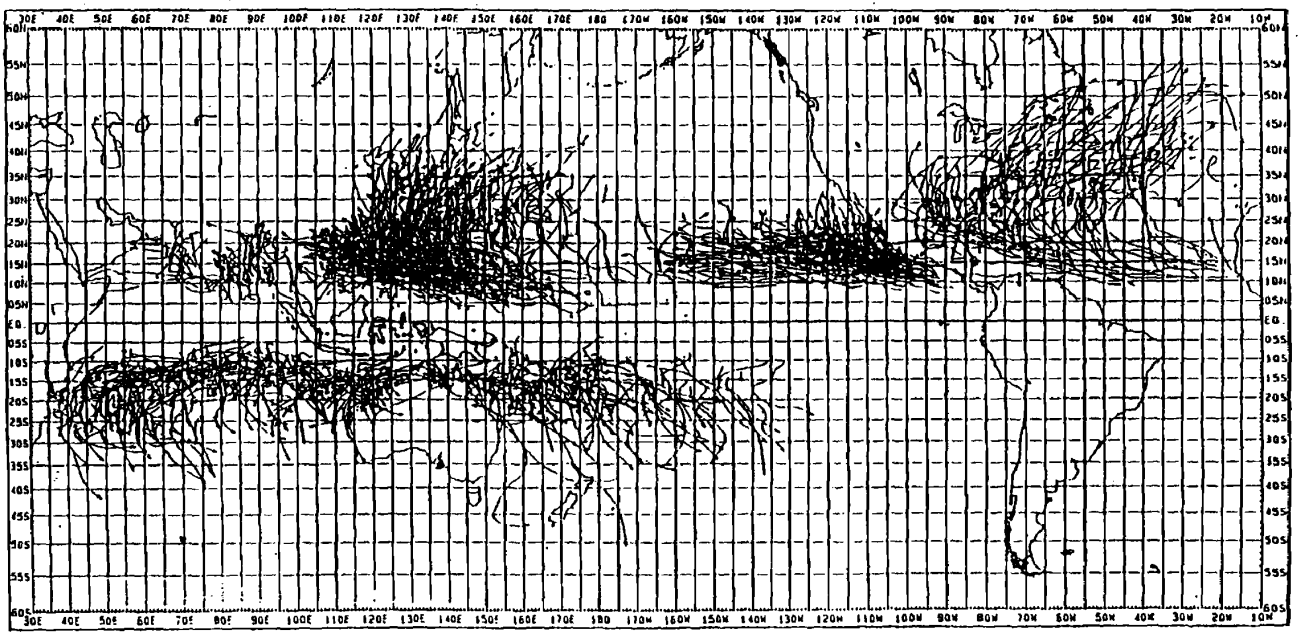


Figure 1 Tracks of all tropical cyclones with winds greater than 39 mph for a 10-year period (Neumann, 1993).

Tropical low	A surface low-pressure system in the tropical latitudes.
Tropical disturbance	A tropical low and an associated cluster of thunderstorms that has, at most, only a weak surface wind circulation.
Tropical depression	A tropical low with a wind circulation of sustained 1-min surface winds of less than 34 knots (kt) [39 miles per hour (mph), 18 meters per second [m/s] circulating around the center of the low]. [A knot (i.e., a nautical mile per hour) equals about 1.15 mph. A nautical mile is the length of 1 min of arc of latitude.]
Tropical storm	A tropical cyclone with maximum sustained surface winds of 34 to less than 64 kt (39 to 74 mph, 18 to 33 m/s).
Hurricane	A tropical cyclone with maximum sustained surface winds of 64 kt (74 mph, 33 m/s) or greater. (In the Pacific Ocean west of the international date line, hurricanes are called typhoons. They are the same phenomenon.)

3 HURRICANES IN NORTH AMERICAN HISTORY

The word *hurricane* derives from the Spanish *huracán*, itself derived from the dialects of indigenous peoples of the Caribbean and Latin America (Dunn and Miller, 1964). 'Hunraken' was the name of the Mayan storm god, and 'Huraken' was the god of thunder and lightning for the Quiche of southern Guatemala (Henry et al., 1994). The Tainos and Caribe tribes of the Caribbean called their God of Evil by the name Huracan. Other indigenous dialects included words such as *aracan*, *urican*, and *hurivanvucan* to refer to "Big Wind." The deification of the hurricane and the connection of indigenous referents with evil and violence is an indication that hurricanes had a significant impact on the lives of many peoples of the Caribbean and Latin America.

The historical record of documented hurricane events begins with the European conquest of North America. Columbus, in his four voyages to North America, experienced direct contact with an Atlantic hurricane only in his fourth voyage. Meteorological historian David Ludlam notes that Columbus' good fortune in his first voyage leads one to wonder "what the course of history in the West Indies might have been if, in the autumn of 1492, a full-blown tropical storm had dashed the frail craft of the Admiral's fleet to the bottom of the sea or flung them shipwreck on some tiny cay" (Ludlam, 1963, p. 1). Others did not experience such good fortune. Shakespeare's play, *The Tempest*, was loosely based on reports of a 1609 hurricane near Bermuda that sunk the vessel *Sea Venture* and stranded the passengers, including John Rolf, future husband of Pocahontas, on the island for 10 months. This storm's movement was among the first successfully anticipated by the colonists. During the course of the storm's trek through the Caribbean, a skipper in the Royal Navy cautioned the British fleet to move out of the storm's path, based on

his experience with the movement of past hurricanes. During the 1700s and 1800s numerous coastal locations were struck by severe hurricanes. Charleston (South Carolina), New Orleans (Louisiana), and Boston (Massachusetts) were particularly hard hit a number of times. In 1772 in the West Indies, teenaged Alexander Hamilton wrote about a hurricane's impact for a local newspaper. His writing caught the attention of the local gentry who then raised money to send him to the mainland colonies to further his education, thus setting the stage for his political career.

Tropical storms were once named after the particular "saint's day" that fell nearest the hurricane event (Tannehill, 1952). For instance, "Hurricane Santa Ana" hit Puerto Rico on 26 July 1825 (see Rodriguez, 1997). Today, tropical cyclones are "named" when they reach tropical storm strength. According to one explanation, this practice dates to the 1950s, following the publication of George R. Stewart's *Storm*, a book that featured a forecaster who named storms (Williams, 1992). Another explanation has the origin of the hurricane naming convention beginning with a military radio operator who, during World War II, ended each hurricane warning singing "Every little breeze seems to whisper Louise," prompting the naming of a particular hurricane Louise (Henry et al., 1994). Whatever the origin, the practice caught on because it proved useful in identifying different storms that existed simultaneously. The personification of the extreme event was also found to be a valuable practice by the various user communities. Until 1979, tropical storms were given only women's names in English. In 1979 forecasters began to use men's, French, and Spanish names as well. The repeating, 6-year list of names assigned to tropical cyclones in the Atlantic was put together by the World Meteorological Organization. It can be found at the National Hurricane Center's website at *http://www.nhc.noaa.gov/names.html*. Hurricanes that cause significant damage or are particularly memorable, such as Andrew (1992), Camille (1969), or Gilbert (1988), are retired and those names are not used again. Table 1 lists retired hurricanes through 1995 and notes death and damages associated with each.

4 GEOGRAPHIC AND SEASONAL DISTRIBUTION: ORIGIN

Typically, in the Atlantic Ocean basin tropical storms and hurricanes develop over warm water between around 10°N to 35°N, generally, during the summer and fall. During an average year about 16 tropical cyclones develop in the eastern Pacific and approximately 10 in the Atlantic including the Gulf of Mexico and Caribbean Sea (Neumann, 1993). During the period of record, tropical cyclones fail to develop south of the equator in the Western Hemisphere east of 130 W because of one or more of the following factors: the relatively cold ocean temperature, typically strong winds in the upper troposphere, or the absence of an initiation area for tropical low-pressure systems with an associated cluster of thunderstorms (Gray, 1968).* Elsewhere these storms develop in the Indian Ocean, western Pacific, and eastern Pacific

*McAdie and Rappaport (1991), however, discussed the formation of a weak tropical cyclone in the south Atlantic west of tropical Africa in 1991.

TABLE 1 "Retired" Atlantic Hurricane Names through 1994

Year	Name	Location	U.S. Costs (1990$) and Total Casualties, etc.
1954	Carol	Louisiana, Mississippi, and Alabama	$2.37 billion, 60 deaths
1954	Hazel	Antilles, North and South Carolina	$144 billion, 1000 deaths
1955	Connie	North Carolina	25 deaths
1955	Diane	Mid-Atlantic and Northeast U.S.	$4.20 billion, 184 deaths
1955	Ione	North Carolina	$444 million
1955	Janet	Lesser Antilles, Belize, and Mexico	538 deaths
1957	Audrey	Louisiana and North Texas	$696 million, 550 deaths
1960	Donna	Bahamas, Florida, and eastern U.S.	$1.82 billion, 364 deaths
1961	Carla	Texas	$1.93 billion, 46 deaths
1963	Flora	Haiti and Cuba	8000 deaths
1964	Cleo	Lesser Antilles, Haiti, Cuba, southeast Florida	$595 million, 213 deaths
1964	Dora	Northeast Florida	$1.16 billion
1964	Hilda	Louisiana	$579 million, 304 deaths
1965	Betsy	Bahamas, southeast Florida, southeast Louisiana	$6.46 billion, 75 deaths
1966	Inez	Lesser Antilles, Hispaniola, Cuba, Florida Keys, Mexico	1000 deaths
1967	Beulah	Antilles, Mexico, South Texas	$844 million; most tornadoes, 115, ever associated with a hurricane
1969	Camille	Louisiana, Mississippi, and Alabama	$5.24 billion, 256 deaths
1970	Celia	South Texas	$1.56 billion
1972	Agnes	Florida, northeast U.S.	$5.24 billion, 122 deaths
1975	Eloise	Antilles, northwest Florida, and Alabama	$1.08 billion
1979	David	Lesser Antilles, Hispaniola, Florida, and eastern U.S.	$487 million, 2000 deaths
1988	Joan	Curacao, Venezuela, Columbia, and Nicaragua	216 deaths; crossed into Pacific and was renamed Miriam
1989	Hugo	Antilles and South Carolina	$7.16 billion, 56 deaths
1990	Diana	Mexico	96 deaths
1990	Klaus	Martinique	
1991	Bob	North Carolina and northeast U.S.	$1.5 billion
1992	Andrew	Bahamas, South Florida, and Louisiana	> $25 billion
1995	Luis	Leeward Islands	$2.5 billion, 16 deaths
1995	Marilyn	Virgin Islands	$1.5 billion, 8 deaths
1995	Opal	Mexico, Florida	$3 billion, 59 deaths
1995	Roxanne	Mexico	$1.5 billion, 14 deaths

After Pielke and Pielke (1997).

TABLE 2 Saffir/Simpson Hurricane Scale

Category	Central Pressure (mbars)	Central Pressure (inches)	Winds (mph)	Surge (ft)	Damage
1	⩾980	⩾28.94	74–95	4–5	Minimal
2	965–979	28.50–28.91	96–110	6–8	Moderate
3	945–964	27.91–28.47	111–130	9–12	Extensive
4	920–944	27.17–27.88	131–155	13–18	Extreme
5	<920	<27.17	>155	>18	Catastrophic

See Pielke and Pielke (1997, p. 17).

north of the equator (Fig. 1). The western north Pacific is the most active area with an annual average of more than 26 tropical cyclones. Globally, there are about 84 tropical cyclones each year with an annual average of 45 that reach hurricane strength (Neumann, 1993).

Hurricanes are classified by their damage potential according to a scale developed in the 1970s by Robert Simpson, a meteorologist and then-director of the National Hurricane Center, and Herbert Saffir, a consulting engineer in Dade County, Florida (Simpson and Riehl, 1981). The Saffir/Simpson scale was developed by the National Weather Service to give public officials information on the magnitude of a storm in progress and is now widely used by producers and users of hurricane forecasts. The scale has five categories, with category 1 representing the least intense hurricane and category 5 the most intense. Table 2 shows the Saffir/Simpson scale and the corresponding criteria for classification.

5 HURRICANE IMPACTS ON OCEAN AND LAND

When a hurricane forms, it poses a significant danger to society. The importance and danger of tropical cyclones differ between land and water. Over the oceans, the human activities and assets at risk are primarily oil rigs, shipping, and air traffic. On land, particularly along the coast, cities, towns, and industrial activities become threatened. Hurricanes also have ecological and geological impacts.

Ocean Impacts

Winds of hurricane speed over the ocean can create monstrous waves. For example, in 1995, the cruise ship *Queen Elizabeth II* was rocked by a 70-ft (21-m) wave caused by distant hurricane Luis. The sea near a hurricane is chaotic, and an extreme hazard to shipping can occur in response to wave motion moving in many directions.

For comparison, strong winds, of course, also occur in winter storms over the open ocean. The risk to shipping and other activities from wave action, however, is generally less serious in such storms for two reasons. First, the wind blows primarily

in one direction in a given sector of a winter storm. Hence the waves move in concert with the wind. A ship can thus orient itself to minimize the effect of the waves. In a hurricane, winds change direction rapidly around the eye. The result is a chaotic sea with swells and waves propagating in a myriad of directions. A ship cannot simply steer into the running sea to reduce its risk since there is no one direction from which the waves come. Large waves also superimpose on top of each other, producing enormous swells.

Land Impacts at the Coast and a Short Distance Inland

At the coast, the major impacts of either a landfalling hurricane or one paralleling the coast are:

- Storm surge
- Winds
- Rainfall
- Tornadoes

Of these weather features, the storm surge has accounted for over 90% of the deaths in a hurricane. In recent years, and particularly in the aftermath of hurricane Andrew, more attention has been paid to the effects of hurricane winds.

Storm Surge

"Storm surge" refers to a rapid rise of sea level that occurs as a storm approaches a coastline. This is in addition to changes in variations in sea level due to tides. Thus, a storm surge causes greatest inundation at high tide. A very strong hurricane may produce a storm surge of 20 ft (6 m), of which about 3 ft (1 m) is due to the lower atmospheric pressure at the center of a hurricane. The remaining storm surge is due to: (i) the piling up of water at the coast, generated by the strong onshore winds and (ii) a decreased ocean depth near the coast, which steepens the surge. A common misconception is that the lower pressure at the center of a storm is the primary cause of the storm surge.

At landfall, storm surge is highest in the front right quadrant of a westward-moving tropical cyclone (in the Northern Hemisphere), where the onshore winds are the strongest. It is also large where ocean bottom bathymetry focuses the wave energy (e.g., as in a narrowing embayment). Peak storm surge from a landfalling cyclone increases with greater wind speeds and the areal extent of the storm's maximum winds, out to about 30 miles (48 km).

Storm surge also occurs when a storm parallels the coast without making landfall. The storm surge will precede the passage of the storm's center when winds blow onshore preceding passage of the eye. Similarly, the surge will lag the storm's center when the hurricane is moving such that onshore winds follow the passage of the eye.

Offshore winds that are associated with a storm can produce a negative surge, as the sea level is lowered by the strong winds blowing out from the coast.

Storm surge is estimated to generally diminish in depth by 1 to 2 ft (0.3 to 0.6 m) for every mile (1.6 km) that it moves inland. Even if the inland elevation were only 4 to 6 ft (1.2 to 1.8 m) above mean sea level, a storm surge of 20 ft (6 m) might typically reach no more than 7 to 10 miles (11 to 16 km) inland. Thus, the most destructive effect of the storm surge hazard is on beaches and offshore islands.

Storm Surge Hazards. A storm surge can be deadly. In 1900, up to 12,000 deaths occurred in Galveston, Texas, primarily as a result of the storm surge that was associated with a Gulf of Mexico hurricane. In 1957, a storm surge was the major cause of death for 390 people in Louisiana. The storm surge, associated with hurricane Audrey, was over 12 ft (3.5 m) in depth and extended as far inland as 25 miles (40 km) in this particularly low-lying region. In September 1928, the waters of Lake Okeechobee, FL driven by hurricane winds, overflowed the banks of the lake and were the main cause of more than 1800 deaths.

Areas to be evacuated due to storm surge in the case of hurricane landfall are determined through a model developed by the National Weather Service (NWS) called SLOSH (sea, lake, and overland surges from hurricanes; Jarvinen and Lawrence, 1985). The SLOSH model is used to define flood-prone areas in 31 "SLOSH basins" along the U.S. Gulf of Mexico and Atlantic coasts (Fig. 2). Determination of storm surge vulnerabilities is the result of an interagency and intergovernmental process funded by the National Oceanic and Atmospheric Administration (NOAA), the Federal Emergency Management Agency (FEMA), Army Corps of Engineers, and various state and local governments (BTFFDR, 1995). From development through application the SLOSH process for a particular location takes about 2 years. Because coastlines are constantly changing due to human and natural forces, the SLOSH process is an ongoing challenge.

Winds

The strong winds of a hurricane can produce considerable structural damage and risk to life from flying debris, even inland from the coast. The damage caused by hurricane Andrew was predominantly due to wind. Although winds reduce after landfall, as the central pressure increases, and the intensity of the storm lessens, destructive winds can still occur far inland.

The damage from winds is proportional to the energy of the airflow, i.e., to the velocity squared; thus, a wind of 100 mph is four times as effective at causing damage as a wind of 50 mph. Maximum gusts, of course, are even stronger than reported sustained winds (which are measured in the United States by averaging wind speed over 1 min). In a hurricane over the open ocean at about 36 ft (11 m) a gust averaged over 2 s is generally about 25% greater than the 1 min average. For flat grassland, the 2-s speed is around 35% larger, while in woods or cities, this measure of gust speeds is 65% greater. Thus a 1 min average wind of 100 mph would be expected to have gusts to 125 mph over the ocean and 165 mph over a forest.

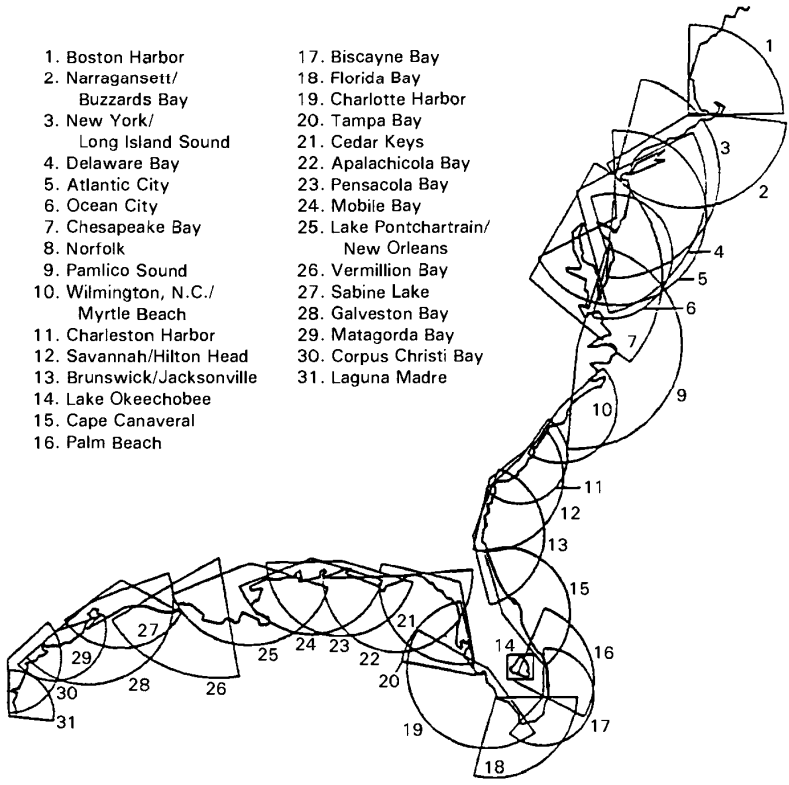


Figure 2 The 31 SLOSH basins along the U.S. Gulf and Atlantic coasts.

Rainfall. Rainfall from hurricanes is beneficial to agriculture, such as the rains from hurricane Dolly (1995) in southern Texas and northeastern Mexico that relieved a drought (Rippey, 1997; cf. Sugg, 1967). Even relatively weak tropical-like disturbances can result in extreme rainfall, as seen, for example, over coastal Texas in September 1979 in which upwards of 19 inches (483 mm) of rain inundated the area over a period of several days (Bosart, 1984). Occasionally, for reasons not completely understood, rainfall is light in the vicinity of hurricanes. Hurricane Inez in 1966, for instance, resulted in only a few drops of rain in Miami for several hours when the center was south and south-southwest of Miami and at its closest point to the city. At the time, Miami was under the storm and, normally, torrential rains would have been expected. As a result of the absence of rain, the strong winds blew salt spray many miles inland, causing severe damage to vegetation from salt accumulation. Homestead Air Force Base, south of Miami and closer to the path

traveled by the hurricane's center, received only 0.62 inches (15.7 mm) of rain during the entire storm.

Tornadoes. Tornadoes are also a threat from tropical cyclones. Much of the damage of Andrew was associated with tornadic vortices whose wind speeds were added onto the large-scale hurricane winds (Black and Wakimoto, 1994). These rapidly rotating small-scale vortices are spawned in squalls, usually in the front right quadrant of the storm with respect to the storm's track.

Wind damage and tornadoes also can occur well inland associated with tropical cyclones. In 1959, hurricane Gracie caused 12 deaths in central Virginia 24 h after landfall on the South Carolina coast. Hurricane Hugo in 1989 caused significant damage in Charlotte, North Carolina, after landfall.

Inland Impacts

Inland, away from the coast, the largest threat to life and property occurs as a result of flash flooding and large-scale riverine flooding from excessive rainfall. Particularly dangerous are tropical cyclones whose rainfall is initially light and benign after landfall only to erupt a couple of days later into torrential downpours when the environment becomes favorable for precipitation of the large quantities of tropical moisture that have moved inland with the storm.

A particularly extreme example of such a system is hurricane Camille of 1969. After killing 139 people along the Gulf coast on August 17, the storm rapidly weakened after moving inland across Mississippi, into Tennessee and Kentucky. There was relatively little concern expressed by the National Weather Service and certainly no hint of the tragedy that was to happen on the night of August 19, 1969, in central Virginia. The 24-h and 12-h precipitation forecasts for the area, for example, indicated that only slightly more than 2 inches (50 mm) were expected. In fact, a deluge occurred in one part of Virginia as the remnants of Camille began to rejuvenate through interaction with a cold front and when the associated moist tropical air was lifted by the mountains. The rainfall of almost 30 inches (760 mm) in 6 h liquefied soils on the mountainous slopes and flooded drainage basins, burying and drowning 109 individuals. As a result of this tragedy, a radar site was installed in southern Virginia. One of the justifications of the new National Weather Service U.S. Doppler radar network (the WSR-D-88 system) is to detect heavy rainfall events.

Such excessive rains well inland from landfalling tropical cyclones should be expected occasionally as occurred over Georgia associated with tropical storm Alberto in 1994. The environment of a storm is a localized region of the atmosphere that is enriched with water vapor, well in excess of even the average tropical environment. After landfall, this rich reservoir of moisture moves inland and can be copiously precipitated when it is lifted through a mechanism such as a mountain barrier and/or ascent over a weather front. Hurricane Agnes in 1972, for instance, produced enormous rainfalls over large areas of the middle Atlantic states because of

strong large-scale atmospheric lifting and the movement of the moist air up and over the Appalachian mountains, resulting in disaster.

Even snowfall has been reported to be associated with the inland portion of a hurricane circulation. In 1963, hurricane Ginny left more than 14 inches (36 cm) of snow in northern Maine as the hurricane moved into Nova Scotia with winds of around 100 mph (45 m/s).

Societal Impacts

When they strike the U.S. coast, hurricanes cost lives and dollars and disrupt communities. Category 3, 4, and 5 storms—intense hurricanes—are responsible for more than 80% of hurricane-related damages. Loss of life, however, occurs from storms of various intensities. Due largely to better warning systems, hurricane-related loss of life has decreased dramatically in the twentieth century (NRC 1989). Yet, in spite of reduced hurricane-related casualties "-a large death toll in a U.S. hurricane is still possible. The decreased death totals in recent years may be as much a result of lack of major hurricanes striking the most vulnerable areas as they are of any fail-proof forecasting, warning, and observing systems" (Hebert et al., 1993, p. 14).

While loss of life has decreased, the economic and social costs of hurricanes are large and rising. A rough calculation shows that annual losses to hurricanes have been in the billions of dollars. In the United States alone, after adjusting for inflation, tropical cyclones were responsible for an annual average of $1.6 billion for the period 1950 to 1989, $2.2 billion over 1950 to 1995, and $6.2 billion over 1989 to 1995 (Hebert et al., 1996). For a comparison, China suffered an average $1.3 billion (unadjusted) in damages related to typhoons over the period 1986 to 1994 (World Meteorological Organization, various years). Significant tropical cyclone damages are also experienced by other countries including those in East Asia (including Japan, China, and Korea) and Southeast Asia, those along the Indian Ocean (including Australia, Madagascar, and the southeast African coast), islands of the Caribbean, and in Central America (including Mexico). While a full accounting of global damages has yet to be documented and made accessible, it is surely in the tens of billions of dollars annually. Other estimates range to $15 billion annually (e.g., Southern, 1992).

Experts have estimated that tropical cyclones result in approximately 12,000 to 23,000 deaths worldwide (Southern, 1992; Smith, 1992; Bryant, 1991). Tropical cyclones have been responsible for a number of the largest losses of life due to a natural disaster. For instance, in April 1991, a cyclone made landfall in Bangladesh resulting in the loss of more than 140,000 lives and disrupting more than 10 million people (and leading to $2 billion in damages; Southern, 1992). A similar storm resulted in the loss of more than 250,000 lives November 1970. China, India, Thailand, and the Philippines have also seen loss of life in the thousands in recent years.

While the hurricane threat to the U.S. Atlantic and Gulf coasts has been widely recognized, it has only been in recent years, following hurricane Andrew, that many

public and private decision makers have sought to better understand the economic and social magnitude of the threat.

One study has sought to "normalize" U.S. hurricane damages to assess the impact that past storms would have had in 1995 (Pielke and Landsea, 1997). The study adjusted past damages to account for changes in population, inflation, and wealth. The study found a total of $366 billion in losses over the period 1925 to 1995, or about $5 billion annually. Interestingly, the normalized data show a trend of *decreasing* losses from the 1940s through the early 1990s, which is contrary to the non-normalized data (Figs. 3 and 4). This highlights the good fortune experienced by the U.S. with respect to hurricane landfalls in recent decades (Landsea et al., 1996).

6 CONCLUSION

Tropical cyclones affect hundreds of millions of people every year around the world. While the rainfall produced by these storms often provides valuable societal and environmental benefits, these storms also have the potential to inflict great harm and suffering. Recent history suggests that communities in the Atlantic basin have been fortunate in recent decades, due to an extended period of relatively fewer hurricanes (Landsea et al., 1996). However, simply because hurricanes have been depressed in

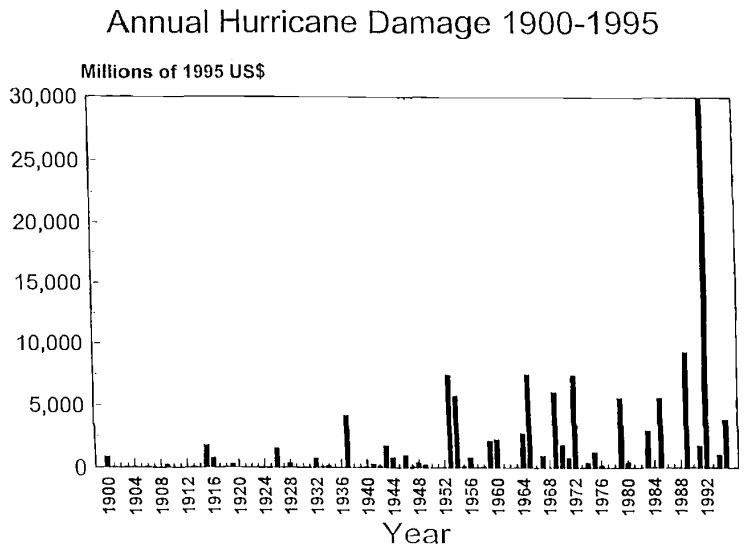


Figure 3 Inflation adjusted hurricane damages of the 20th century. (Pielke and Landsea, 1997).

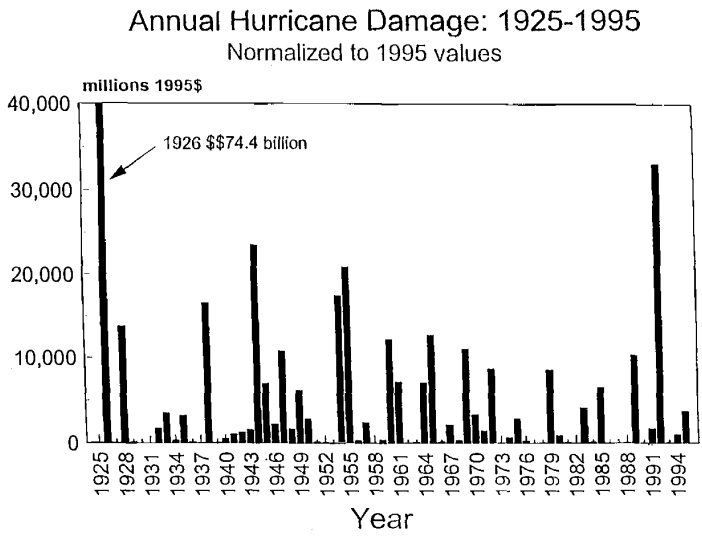


Figure 4 Hurricane damages adjusted for inflation, wealth, and population 1925 to 1995 (Pielke and Landsea, 1997).

recent decades does not eliminate the possibility of large impacts, as shown by hurricane Andrew, which occurred during the quietest 4-year period of hurricane activity since 1950. The $30 billion hurricane Andrew was the costliest tropical cyclone ever (Landsea et al., 1996; Pielke and Pielke, 1997).

Tropical cyclones occur every year around the world. In this most basic sense, they are "normal" climatological events on planet Earth. But from a human perspective, even a weak tropical cyclone can be an "extreme" occurrence. The challenge of effectively reducing societal vulnerability to hurricanes is made more difficult by the relative infrequency with which storms affect particular communities. Consider that the last major hurricane to strike Dade County, Florida, prior to Andrew was in 1950! Thus, one important step any decision maker should take is to understand the risks and potential consequences of choices made in tropical cyclone-prone regions. Damaging losses associated with tropical cyclones can never be eliminated, but with close attention to those factors that increase our vulnerability—where we live, how we live, etc.—we can hope to live in greater harmony with one of nature's most powerful forces.

REFERENCES

Ager, D., *The New Catastrophism*. Cambridge University Press, Cambridge, 1993.

Anthes, R. A., *Tropical Cyclones: Their Evolution, Structure and Effects*, American Meteorological Society, Boston, MA, 1982.

Bipartisan Task Force on Funding Disaster Relief BTFFDR. *Federal Disaster Assistance: Report of the Senate Task Force on Funding Disaster Relief*, No. 104-4, U.S. Government Printing Office, Washington, DC, 1995.

Black, P. G., and R. M. Wakimoto, Damage survey of hurricane Andrew and its relationship to the eyewall, *Bull. Am. Meteor. Soc.*, *75*, 189–200, 1994.

Bosart, L. F., The Texas coastal rainstorm of 17–21 September 1979: An example of synoptic-mesoscale interaction, *Monthly Weather Rev.*, *112*, 1108–1133, 1984.

Bryant, E. A., *Natural Hazards.* Cambridge University Press, Cambridge, 1991.

Dunn, G. E., and B. I. Miller, *Atlantic Hurricanes*, Louisiana State University Press., Baton Rouge, LA, 1964.

Elsberry, R. L., W. M. Frank, G. J. Holland, J. D. Jarrell, and R. L. Southern, A global view of tropical cyclones, based largely on materials prepared for the International Workshop on Tropical Cyclones, Bangkok, Thailand, November 25–December 5, 1985, Office of Naval Research, Marine Meteorology Program, Robert F. Abbey, Director, 1987.

Gibbs, W. J., Defining climate. *WMO Bull.*, *36*, 290–296, 1987.

Gray, W. M., A global view of the origin of tropical disturbance and storms, *Monthly Weather Rev.*, **96**, 669–700, 1968.

Hebert, P. J., J. D. Jarrell, and M. Mayfield, *The Deadliest Costliest, and Most Intense United States Hurricanes of This Century*, NOAA NWS NHC-31, 1993.

Hebert, P. J., J. D. Jarrell, and M. Mayfield, *The Deadliest, Costliest, and Most Intense United States Hurricanes of This Century (and Other Frequently Requested Hurricane Facts)*, NOAA Technical Memorandum NWS TPC-1, National Hurricane Center, Miami, FL, February 1996.

Henry, J. A., K. M. Portier, and J. Coyne, *The Climate and Weather of Florida*, Pineapple Press, Sarasota, FL, 1994.

Jarvinen, B. R., and M. B. Lawrence, An evaluation of the SLOSH storm-surge mode, *Bull. Am. Meteor. Soc.*, *66*, 1408–1411, 1985.

Katz, R. W., and B. G. Brown, Extreme events in a changing climate: Variability is more important than averages, *Climatic Change*, *21*, 289–302, 1992.

Katz, R. W., Towards a statistical paradigm for climate change, *Climate Res.* **2**, 167–175, 1993.

Landsea, C. W., N. Nicholls, W. M. Gray and L. A. Avila, Quiet early 1990s continues trend of fewer intense Atlantic hurricanes, *Geophys. Res. Lett.*, *23*, 1697–1700, 1996.

Ludlam, D. M., *Early American Hurricanes*: 1492–1870, American Meteorological Society, Boston, MA, 1963.

McAdie, C. J., and E. N. Rappaport, *Diagnostic Report of the National Hurricane Center*, Vol. 4, No. 1, NOAA, National Hurricane Center, Coral Gables, FL, 1991.

Neumann, C. J., Global overview, in *Global Guide to Tropical Cyclone Forecasting*, World Meteorological Organization (WMO) Technical Document, WMO/TD NO. 560, Tropical Cyclone Programme, Report No. TCP-31, WMO, Geneva, Switzerland, Chapter 1, 1993.

Neumann, C. J., B. R. Jarvinen, and A. C. Pike, *Tropical Cyclones of the North Atlantic Ocean, 18711986* , 3rd rev., NOAA Historical Climatology Series 6-2, NCDC; Asheville, NC, 1987.

National Research Council (NRC), *Opportunities to Improve Marine Forecasting*, National Academy Press, Washington, DC, 1989.

Pielke, Jr., R. A., and C. W. Landsea, Normalized hurricane damages in the United States 1929–1995, *Weather Forecast.*, 1997.

Pielke, Jr. R. A., and R. A. Pielke, *Hurricanes: Their Nature and Impacts on Society*, J Wiley, New York, 1997.

Rippey, B., Weatherwatch—August 1996, *Weatherwise*, *49*, 51–53, 1997.

Rodriquez, H., A socioeconomic analysis of hurricanes in Puerto Rico: An overview of disaster mitigation and preparedness, in H. F. Diaz and R. S. Pulwarty (Eds.), *Hurricanes*, Springer-Verlag, Berlin, 1997, pp. 121–143.

Simpson, R. H., and H. Riehl, *The Hurricane and Its Impact*, Louisiana State University Press, Baton Rouge, LA, 1981.

Smith, K., *Environmental Hazards: Assessing Risks and Reducing Disaster*, Routledge, London, 1992.

Southern, R. L., Savage impact of recent catastrophic tropical cyclones emphasizes urgent need to enhance warning/response and mitigation systems in the Asia/Pacific region, unpublished.

Sugg, A. L., Economic aspects of hurricanes, *Monthly Weather Rev.*, *95*, 143–146, 1967.

Tannehill, I. R., *Hurricanes: Their Nature and History*, 8th ed., Princeton University Press, Princeton, NJ, 1952.

Williams, J., *The Weather Book*. Vintage Books, New York, 1992.

CHAPTER 43

EL NIÑO IN AUSTRALIA

NEVILLE NICHOLLS

1 INTRODUCTION

Before the 1972–1973 El Niño episode, understanding of the impacts of the El Niño–Southern Oscillation (ENSO) on Australia was limited. Studies in the 1970s and 1980s documented its effects, but the 1982–1983 event still caught the country by surprise. By the El Niño events of the early 1990s, a routine seasonal climate prediction service, based on the earlier work on the ENSO, had been established.

2 EL NIÑO–SOUTHERN OSCILLATION EFFECT ON AUSTRALIAN CLIMATE

Australian droughts generally accompany El Niño episodes (e.g., Allan, 1991). Figure 1 illustrates the relationship between widespread Australian drought and low values of the Southern Oscillation Index (the SOI, a simple measure of the ENSO, is the standardized difference in surface atmospheric pressure between Tahiti and Darwin), by comparing time series of the percentage of Australia with annual rainfall in the lowest decile with annual averages of the SOI. The figure also indicates that years with little of the country in drought tend to have large positive SOI values, ie., La Niña episodes.

Figure 1 only uses data from 1950, for clarity. The relationship between the SOI and drought, however, is evident in data throughout the twentieth century. Prior to the late nineteenth century there are insufficient data to allow a strict, quantitative comparison of widespread Australian droughts with the ENSO. Nicholls (1988) examined reports of the governors of the colony of New South Wales to the colonial secretary of the British Government in London for references to drought in the early

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

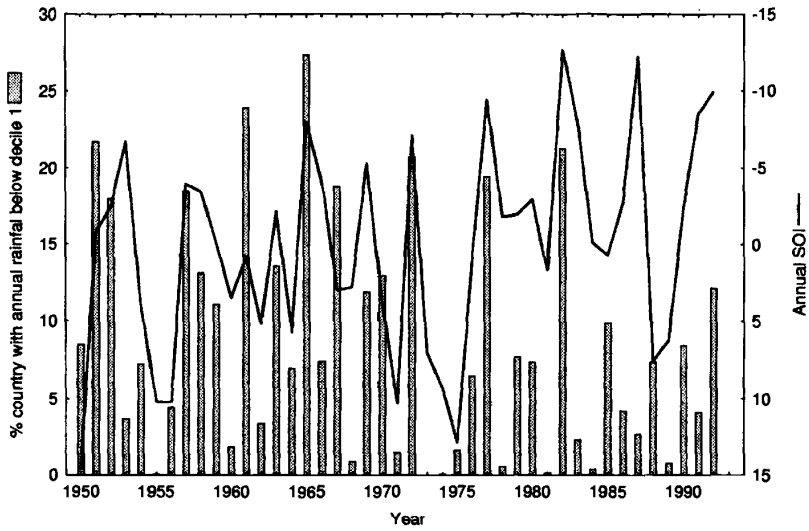


Figure 1 Annual mean SOI (full line) and the percentage of Australia with annual rainfall below the first decile, i.e., in drought, (broken line).

years of the colony and found that the coincidence of El Niño events and Australian droughts has existed from, at least, the start of European colonization in 1788.

The ENSO also enhances Australian rainfall variability, as it does wherever it impacts on climate (Nicholls et al., 1997). Also, many Australian droughts tend to last about a year because El Niño and La Niña episodes both tend to last about 12 months and this sets the time scale of Australian rainfall fluctuations (Nicholls, 1991). The link with the ENSO is most consistent with east and north Australian rainfall (e.g., Pittock, 1975; McBride and Nicholls, 1983; Ropelewski and Halpert, 1987, 1989).

3 DISCOVERY OF EFFECT OF EL NIÑO–SOUTHERN OSCILLATION ON AUSTRALIA

India suffered a severe drought and famine during 1877. Sir Henry Blanford, the director of the Indian Meteorological Service, noted the very high atmospheric pressures over Asia at the time and requested pressure information from other meteorologists around the world. Sir Charles Todd, the South Australian government observer noted that pressures were also high during 1877 over Australia, and much of the country suffered from drought that year. Todd compared earlier droughts and concluded that Indian and Australian droughts usually coincided. This observation has since been confirmed (e.g., Williams et al., 1986) and forms part of the suite of climate linkages we now call the Southern Oscillation (SO).

When Sir Gilbert Walker named and documented the SO in the early decades of the twentieth century, its close relationship with Australian rainfall quickly became apparent (e.g., Bliss and Walker, 1932). Walker's work suggested that north Australian summer rainfall could be predicted with an index of the SO. Quayle (1910, 1929) suggested that rainfall farther south could be predicted in the same way. After that, a trickle of studies discussed the relationship between the SO and Australian climate, up to the mid-1970s, when the worldwide attention on El Niño led to a resurgence of interest among Australian meteorologists.

By the early- 1980s attention had turned to the possible use of the ENSO in prediction. Work on the physical cause of the phenomenon had commenced, and several studies describing patterns and relationships between the ENSO, sea surface temperature, and Australian climate had been published (e.g., Pittock, 1975; Streten, 1981; Coughlan, 1979). Some of the lag relationships suggested by Quayle and others had been validated and extended using new data (Nicholls and Woodcock, 1981; McBride and Nicholls, 1983). New relationships indicating that seasonal temperature, wet-season onset, and even seasonal tropical cyclone activity also were predictable, through the ENSO, had been uncovered (Nicholls, 1978, 1979; Nicholls et al., 1982). The recognition in mid-1982 that a major El Niño episode was under way led to cautious statements regarding possible implications for Australian rainfall through the remainder of 1982, based on this work (Nicholls, 1983). The Bureau's National Climate Centre began preparing and issuing regular monthly "Seasonal Climate Outlooks" in 1989, based on the SOI. These provide forecasts of 3-month rainfall anomalies, across the country.

Variables other than seasonal rainfall appear to be predictable through the use of the ENSO. For instance, Stone et al. (1996) suggest that seasonal frost forecasts could be feasible in eastern Australia. Nicholls and Kariko (1993) and Suppiah and Hennessy (1996) found that rainfall events and intensity were related to the ENSO. Whetton et al. (1990) and Allan et al. (1996) documented relationships of the ENSO with streamflow variations.

4 ECOLOGICAL IMPACTS OF EL NIÑO–SOUTHERN OSCILLATION

The widespread effect of the ENSO on Australian climate variations suggests that there should be strong responses to the phenomenon in Australian biota, including crops. There is ample evidence that the high variability the ENSO imposes on Australian climate impacts the wildlife and vegetation; populations of many Australian animals are maintained at levels well below the carrying capacity of the good years. Populations typically increase dramatically during periods of good rainfall and fall during the frequent droughts. Some adaptations to variable precipitation observed in the Australian native biota are described by Nicholls (1989, 1991). Each adaptation allows opportunistic use of good conditions, thereby producing rapid increases in populations when a drought breaks.

Red Kangaroo

Australia's largest herbivore, the red kangaroo, inhabits the open arid and semiarid plains that cover most of the continent. It shows no seasonal pattern of reproduction but breeds opportunistically in response to good conditions by producing young in rapid succession. Under prolonged drought the kangaroos stop breeding. Drought-breaking rains trigger an immediate hormonal response. The females return rapidly to breeding and may be found with young in the pouch after 60 days. In favorable environmental conditions females become sexually mature when 15 to 20 months old. Drought delays the onset of sexual maturity and after 2 years of drought a population may include females aged 3 years or more that have never produced young. After rain these animals come into breeding condition almost immediately. The life-history strategy of the red kangaroo is clearly adapted to highly variable rainfall.

Green Turtles

The number of green turtles observed nesting around northern Australia varies widely from year to year, and these interannual fluctuations are in phase at widely separated rookeries, with large numbers of turtles breeding 2 years after major El Niño episodes (Limpus and Nicholls, 1988). Preparation for breeding commences well over a year before oviposition. Atmospheric or oceanic anomalies associated with El Niño (perhaps increased availability of food due to the reduced number of tropical cyclones during El Niño) triggers the turtles to commence breeding. The relationship with El Niño provides a means for predicting, a long way in advance, the approximate numbers of turtles breeding. Such a prediction is potentially useful in sea turtle management in areas where eggs, courting turtles, or nesting females are harvested.

Australian Encephalitis

Australian encephalitis (AE) is a severe, often fatal, viral illness transmitted to humans by mosquitoes and influenced by the ENSO (Nicholls, 1986). Since 1917 when the clinical symptoms of the illness were first diagnosed, there have been only 7 years when cases of AE were observed in southeastern Australia. Cases occur between January and April and follow widespread flooding over several seasons. Flooding leads to increased mosquito numbers by increasing the numbers of breeding sites and host populations (birds, marsupials). The probability that AE will be diagnosed in southeastern Australia is predictable from the SOI in the previous spring (September–November). The relationship is sufficiently strong to allow health authorities to increase surveillance and prophylactic action, in years when the SOI, during spring, is very high.

Banana Prawns

The prawn season in the Gulf of Carpentaria extends from March to June. Prawn catch is related to the amount of rainfall. Mature prawns require a saline environment. With the advent of heavy wet-season rainfall, salinity in the rivers where the prawns breed is lowered, and they are forced to migrate offshore. Once offshore the prawns may be harvested. If wet-season rainfall is very low, then fewer prawns leave the rivers, leading to a low catch. The relationship between the SOI and rainfall in this region means that there is a significant relationship between the SOI in November and the subsequent prawn catch (Love, 1987). Low values of the SOI lead to a lower than normal catch.

Waterfowl

The number of ducks shot on the opening day of annual waterfowl season in southeast Australia is correlated with the SOI some 2 year prior to the season (Norman and Nicholls, 1991). Apparently, heavy rains (which often follow an El Niño–related drought) result in widespread floods. These fill ephemeral wetlands, leading to enhanced waterfowl breeding (few species breed on permanent waters). In the following year, as the post-El Niño floods recede, the waterfowl congregate in the permanent wetlands, thereby leading to inflated numbers. An increased harvest ensues in the next open season. So, an El Niño, with low SOI, tends to be followed about 2 years later by an increased duck harvest.

Australian Birds

Some other adaptations of Australian birds that can be linked to the unpredictable environment (in turn caused by the ENSO) are nomadism, irregular and seasonal breeding initiated by sudden falls of rain, variations in clutch and multiple broods dependent on the rainfall, precocious breeding, and the habit of older offspring of assisting the breeding male in raising siblings of subsequent broods. All these behavior patterns contribute to an opportunistic life-history strategy. Population increases can be very rapid after drought, just as with the red kangaroo.

Vegetation

Vegetation also is linked to the high variability caused by ENSO. The following are just some of the characteristics of Australian vegetation that may be, at least in part, attributable to the ENSO influence on the climate: absence of succulents, establishment dependent on extended wet periods, drought tolerance/avoidance, diverse life-history strategies, and fire resistance/dependence (Nicholls, 1991).

These are just a few examples of the many and varied adaptations of Australian biota to the highly variable rainfall found in some members of most groups of Australian plants and animals. There appears to be a consensus among ecologists that much of the Australian flora and fauna is adapted to a highly variable rainfall

that, in turn, is caused by the ENSO, and that this adaptation is more complete than in other areas of the globe.

5 EL NIÑO–SOUTHERN OSCILLATION AND VEGETATION CHANGES

Since the native Australian vegetation was adapted to the climate rhythms and variability induced by the ENSO, it is not surprising that the introduction of plants and animals not so adapted led to rapid changes in vegetation (Nicholls, 1991). The best known of these changes is probably the area now known as the Pilliga Scrub in northern New South Wales (Rolls, 1981; Austin and Williams, 1988). Much of this area of 400,000 ha was open grassy country with only about eight large trees per hectare when Europeans arrived in the 1830s. Frequent burning by Aboriginals, and grazing by indigenous marsupials, restricted the opportunities for trees and shrubs to establish. Fire germinated the seed of the trees and shrubs, but rat kangaroos ate many of the resulting seedlings before they could establish.

The introduction of sheep reduced the numbers of rat kangaroos, by destroying their cover and their food. A severe drought during the major El Niño of 1877–1978 further reduced the numbers of indigenous marsupials. The following year, a major La Niña event, was very wet. The few large trees seeded well and when stock owners burnt to destroy grasses with seeds that got into their sheep's wool, seedlings came up thickly, unhindered by the grasses that would usually compete with them for space. This time there were no rat kangaroos to eat the seedlings either and the trees grew unchecked.

Over the next decade there were several further periods of establishment, again synchronized with El Niño–La Niña oscillations. The European rabbit, also an enthusiastic eater of seedlings, arrived in the area in the late 1880s and prevented further establishment until myxomatosis in 1951 reduced the rabbit population. The first successful release of myxomatosis occurred in 1950. Earlier releases of the disease had not led to widespread establishment. The extensive rains and flooding in 1950, associated with a major La Niña, contributed to the successful establishment of the disease by providing ideal breeding conditions for the insects that spread it.

In 1917 the Forestry Commission stopped burning in the Pilliga and by 1950 large amounts of forest litter had accumulated. So had decades of seed production. The forest dried in El Niño event of 1951, following good growth during La Niña of 1950, and a major fire started in November 1951. In the absence of rat kangaroos and rabbits, the new growth induced by the fire had nothing to stop it.

In less than a century Europeans had unintentionally transformed the area from grazing land into the dense Pilliga Scrub supporting sustained timber harvesting. The ENSO phenomenon played a critical role in this transformation. McKeon et al. (1990) cite other examples where the extreme climate events associated with both extremes of the ENSO resulted in major long-term vegetation degradation. In western Queensland there was a rapid increase in the sheep population during the above-average rainfall years of the early 1890s. Major El Niño events between 1899

and 1902 resulted in very low rainfall and a rapid drop in animal numbers. Heavy utilization of edible grasses and shrubs during this drought led to a spread of inedible plants and carrying capacities seem to have been permanently reduced. In the subtropical grasslands of southern coastal Queensland, rapid change in species composition to bunch spear grass appears to have resulted from overgrazing with sheep during El Niño–related drought of 1881–1882. More recently, low beef prices in the mid-1970s led to increased stocking rates in Queensland. These years were wet, the result of the 1973–1975 La Niña, but attempts to maintain the high stocking rates into the 1980s with their drier, El Niño conditions have led to pasture degradation, species changes, and soil erosion.

6 IMPACTS OF EL NIÑO–SOUTHERN OSCILLATION ON AUSTRALIAN CROPS

Not surprisingly, given its effects on Australian climate, the ENSO has a major impact on crop yields. Figure 2 shows time series of wheat yields, averaged across Australia, and the SOI. The year-to-year differences in the two variables are plotted, to remove the effects of trends and changes such as the introduction of new cultivars. The relationship is clear—negative values of the SOI lead to widespread drought (Fig. 1), which leads to low crop yields (Nicholls, 1985).

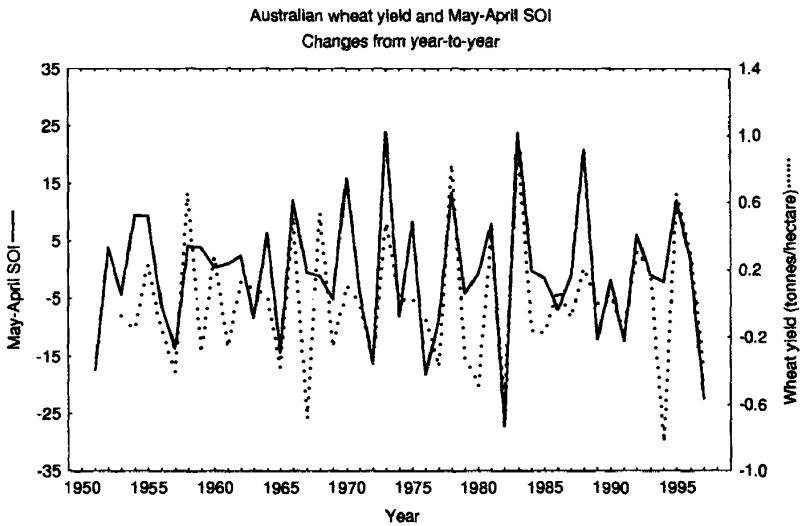


Figure 2 Annual mean SOI (full line) and the yield (tonnes per hectare) of wheat, averaged across the country. Differences in SOI and yield plotted, to remove long-term variations such as the effects of changes in cultivars.

Rimmington and Nicholls (1993) demonstrated that wheat yields in all states were correlated with values of the SOI from before and near the sowing date, which therefore can provide skillful yield forecasts of Australia's major crop. These forecasts would be available several months before harvest starts, require little data, and are quick and simple to prepare. Strong negative relationships also exist with the SOI in the year before the crop is planted, i.e., an El Niño episode is often followed by good crops the following year. This partly reflects the biennial nature of the ENSO, but may also reflect a tendency for a drop in pests in the droughts associated with negative SOI values. This would amplify any response of the crops to good rains in the following year.

Hammer et al. (1996) examined the value of ENSO-based forecasting methodologies to wheat crop management in northern Australia, by examining decisions on nitrogen fertilizer and cultivar maturity using simulation analyses of specific production scenarios. The average profit and risk of making a loss were calculated for the possible range of fixed (i.e., the same each year) and tactical (i.e., varying depending on the ENSO-based seasonal forecast) strategies. Significant increases in profit (up to 20%) and/or reduction in risk (up to 35%) of making a loss were associated with the tactical (forecast-based) strategies. The skill in seasonal rainfall and frost predictions, based on the ENSO, generated the value from using tactical management. This study demonstrated that the skill obtainable in Australia was sufficient to justify, on economic grounds, their use in crop management. Presumably, these forecasts could also be useful in drought management decision making, for instance, in determination of appropriate stocking rates on pastoral properties (McKeon et al., 1990).

REFERENCES

Allan, R. J., Australasia, in M. Glantz, R. Katz, and N. Nicholls (Eds.), *Teleconnections Linking Worldwide Climate Anomalies*, Cambridge University Press, Cambridge, 1991, pp. 73–120.

Allan, R. J., G. S. Beard, A. Close, A. L. Herczeg, P. D. Jones, and H. J. Simpson, *Mean Sea Level Pressure Indices of the El Niño–Southern Oscillation: Relevance to Stream Discharge in South-eastern Australia*, Divisional Report 96/1, *CSIRO Division of Water Resources*, Canberra, Australia, 1996.

Austin, M. P., and O. B. Williams, Influence of climate and community composition on the population demography of pasture species in semi-arid Australia, *Vegetatio*, *77*, 43–49, 1988.

Bliss, E. W., and G. T. Walker, World weather V, *Mem. R. Meteorolo. Soc.*, *4*, 52–84, 1932.

Coughlan, M. J., Recent variations in annual-mean maximum temperatures over Australia, *Q. J. R. Meteorol. Soc.*, *105*, 707–719, 1979.

Hammer, G. L., D. P. Holzworth, and R. Stone, The value of skill in seasonal climate forecasting to wheat crop management in a region with high climatic variability, *Aust. J. Agric. Res.*, *47*, 717–737, 1996.

Limpus, C. J., and N. Nicholls, The Southern Oscillation regulates the annual numbers of green turtles (*Chelonia mydas*) breeding around northern Australia, *Austral. J. Wildlife Res.*, *15*, 157–161, 1988.

Love, G., Banana prawns and the Southern Oscillation Index, *Austral. Meteorol. Mag.*, *35*, 47–49, 1987.

McBride, J. L., and N. Nicholls, Seasonal relationships between Australian rainfall and the Southern Oscillation, *Monthly Weather Rev.*, *111*, 1998–2004, 1983.

McKeon, G. M., K. A. Day, S. M. Howden, J. J. Mott, D. M. Orr, W. J. Scattini, and E. J. Weston, Management of pastoral production in northern Australian savannas, *J. Biogeog.*, *17*, 355–372, 1990.

Nicholls, N., A possible method for predicting seasonal tropical cyclone activity in the Australian region, *Monthly Weather Rev.*, *107*, 1221–1224, 1978.

Nicholls, N., A simple air-sea interaction model, *Q. J. R. Meteorol. Soc.*, *105*, 93–105, 1979.

Nicholls, N., Predictability of the 1982 Australian drought, *Search*, *14*, 154–155, 1983.

Nicholls, N., Impact of the Southern Oscillation on Australian crops, *J. Climatol.*, *5*, 553–560, 1985.

Nicholls, N., A method for predicting Murray Valley Encephalitis in southeast Australia using the Southern Oscillation, *Austral. J. Exper. Biol. Med. Sci.*, *64*, 587–594, 1986.

Nicholls, N., More on early ENSOs: Evidence from Australian documentary sources, *Bull. Am. Meteorol. Soc.*, *69*, 4–6, 1988.

Nicholls, N., How old is ENSO? *Climatic Change*, *14*, 111–115, 1989.

Nicholls, N., The El Niño–Southern Oscillation and Australian vegetation, *Vegetatio*, *91*, 23–36, 1991.

Nicholls, N., and A. P. Kariko, East Australian rainfall events: Interannual variations, trends, and relationships with the Southern Oscillation, *J. Climate*, *6*, 1141–1152, 1993.

Nicholls, N., and F. Woodcock, Verification of an empirical long-range weather forecasting technique, *Q. J. R. Meteorol. Soc.*, *107*, 973–976, 1981.

Nicholls, N., J. L. McBride, and R. J. Ormerod, On predicting the onset of the Australian wet season at Darwin, *Monthly Weather Rev.*, *110*, 14–17, 1982.

Nicholls, N., W. Drosdowsky, and B. Lavery, Australian rainfall variability and change, *Weather*, 1997.

Norman, F. I., and N. Nicholls, The Southern oscillation and variations in waterfowl abundance in southeastern Australia, *Austral. J. Ecol.*, *16*, 485–490, 1991.

Pittock, A. B., Climatic change and the patterns of variation in Australian rainfall, *Search*, *6*, 498–504, 1975.

Quayle, E. T., *On the Possibility of Forecasting the Approximate Winter Rainfall for Northern Victoria*, Bulletin No. 5, Commonwealth Bureau of Meteorology, Melbourne, March 1910.

Quayle, E. T., Long range rainfall forecasting from tropical (Darwin) air pressures, *Proc. R. Soc. Victoria*, *41*, 160–164, 1929.

Rimmington, G. M., and N. Nicholls, *Austral. J. Agric. Res.*, *44*, 625–632, 1993.

Rolls, E. C., *A Million Wild Acres*, Nelson, Melbourne, 1981.

Ropelewski, C. F., and M. S. Halpert, Global and regional scale precipitation patterns associated with the El Niño–Southern Oscillation, *Monthly Weather Rev.*, *115*, 1606–1626, 1987.

Ropelewski, C. F., and M. S. Halpert, Precipitation patterns associated with the high index phase of the Southern Oscillation, *J. Climate*, *2*, 268–284, 1989.

Stone, R., N. Nicholls, and G. Hammer, Frost in northeast Australia: Trends and influences of phases of the Southern Oscillation, *J. Climate*, *9*, 1896–1909, 1996.

Streten, N. A., Southern Hemisphere sea surface temperature variability and apparent associations with Australian rainfall, *J. Geophys. Res.*, *86*, 485–497, 1981.

Suppiah, R., and K. J. Hennessy, Trends in the intensity and frequency of heavy rainfall in tropical Australia and links with the Southern Oscillation, *Austral. Meteorol. Mag.*, *45*, 1–18, 1996.

Whetton, P., D. Adamson, and M. Williams, Rainfall and river flow variability in Africa, Australia and East Asia linked to El Niño–Southern Oscillation events, *Geol. Soc. Austral. Symp. Proc.*, *1*, 71–82, 1990.

Williams, M. A. J., D. A. Adamson, and J. T. Baxter, Late Quaternary environments in the Nile and Darling basins, *Austral. Geogr. Stud.*, *24*, 128–144, 1986.

CHAPTER 44

BIOLOGICAL AND SOCIETAL IMPACTS OF CLIMATE VARIABILITY: AN EXAMPLE FROM PERUVIAN FISHERIES

KENNETH BROAD

1 INTRODUCTION

The collapse of the massive industrial anchovy fishery in Peru in 1973 brought El Niño to the world's attention. However, small-scale (artisanal) fishermen from northern Peru and southern Ecuador were aware of this phenomenon long before this widely publicized event. They realized at least a century ago that every few years, around Christmas time, a warm water current appeared close to their desert shores; they named this current "El Niño", or "the boy child," after the baby Jesus. The first time the term El Niño appeared was in a Peruvian newspaper in 1891, as a result of what we now know was a very strong event. Historical reconstruction (Quinn et al., 1987) of data from ship's logs, fish landings, bird populations, crop yields, among other indicators, have documented El Niño events back several hundred years, and others have argued, using proxy evidence, that El Niño has recurred over many millennium (Rodbell et al., 1999).

After intensive studies and data accumulation, it was discovered that El Niño was not only an oceanographic phenomenon related to the western coast of South America but a complex interaction between the ocean (sea surface temperature changes in the equatorial Pacific) and atmosphere (sea-level pressure changes in the western equatorial Pacific)—hence called the El Niño–Southern Oscillation (ENSO). ENSO is now known to impact climate patterns around the globe. Nonetheless, the Peruvian coast remains one of the areas most consistently and directly impacted by this

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

recurrent event. ENSO also has a less common, cold phase, sometimes referred to as La Niña. Except for scientists who study ENSO, most Peruvians do not differentiate La Niña from "normal" conditions. (This chapter uses the term ENSO to refer to El Niño or ENSO warm events.)

While most associate ENSO with negative impacts on flora and fauna, this cyclical climate event, which thus far has presented itself differently with each occurrence in terms of its strength, duration, and intensity, has a range of both negative and positive ecological impacts. These environmental changes trigger varied socioeconomic and political reactions that in turn may alter aspects of that society as a whole over time. In addition, society undergoes change. A changed society will then react differently to the next ENSO event. This makes planning for mid to long-term climate variability a challenging task for governments and individuals. In the last decade and a half, much effort has been put into better understanding and predicting regional climate variability associated with ENSO events, and forecasts are now being used by governments and individuals in many parts of the world in their planning efforts.

This section discusses some of the key impacts of ENSO on the fisheries sector of Peru, including a brief description of what ENSO is and the evolution of scientific interest in this phenomenon; a brief overview of the Peruvian fisheries sector, with examples of impacts from the 1997–1998 ENSO event; and policy implications of ENSO-related climate forecasts.

2 WHAT IS ENSO?

The El Niño–Southern Oscillation (ENSO) is a coupled atmospheric–oceanic phenomenon that has global manifestations and recurs approximately every 2 to 10 years. The atmospheric component of ENSO is the Southern Oscillation, an interannual seesawing of sea-level atmospheric pressure anomalies between northern Australia (Darwin) and the southeast Pacific (near Tahiti). There is both a "warm phase" (El Niño) and a "cold phase" (La Niña). The warm phase involves an extensive warming of the upper ocean along the central and eastern equatorial Pacific and a depression of the thermocline (the boundary that separates the warmer mixed upper layer of the ocean from the cold abyss) in the eastern tropical Pacific. The cold phase involves a cooling of the upper ocean and a rise of the thermocline toward the ocean's surface in the eastern tropical Pacific.

During an ENSO warm phase, as the thermocline deepens, wind-driven coastal upwelling off the shores of South America carries warmer water than usual to the surface. Coastal seasurface temperature anomalies as high as 10°C have been recorded off Peru (Sharp and McLain, 1993).

The Southern Oscillation leads to a cyclic increase and decrease in the strength of the Southern Hemisphere (southeast) trade winds. These winds are strongest during the oceanic cold phase, when sea-level atmospheric pressure in northern Australia, normally low, is anomalously lower, while that in the southeast Pacific, normally high, is anomalously higher. During the oceanic warm phases, when the sea-level

pressure in northern Australia is anomalously high and that in the southeast Pacific is anomalously low, the trade winds weaken, and in extreme cases even blow westerly. There is positive interaction (e.g., feedback) between the ocean and atmosphere, as increased sea surface temperatures increase atmospheric pressure (Philander, 1990). For a comprehensive overview of the observations and mechanisms of the 1997–1998 ENSO see McPhaden (1999).

Dramatic shifts of flora and fauna in waters off southern Colombia and Ecuador to Peru and northern Chile are linked to ENSO events (Arntz et al., 1985). In severe events, the increased ocean temperatures and reduced concentrations of phytoplankton negatively impact some pelagic species, such as the commercially important anchovy and sardines (see Figure 1). Tropical species of fish, however, may extend their ranges to the south and closer to shore, as warmer waters appear along the Peruvian coast. For an overview of the 1997 to 1998 event on biogeochemical cycles and on the use of satellite technology during this event, see Chavez et al. (1998), McPhaden (1999), and Carr and Broad (2000).

The workings of ENSO were first investigated in the mid-1960s by a researcher named Jacob Bjerknes (1966). Bjerknes linked the oceanic process off the Peruvian coast with the seesaw in atmospheric pressure between the western and central equatorial Pacific (i.e., the Southern Oscillation). A growing interest in the possible global connection of climatic events led to the establishment of the World Climate

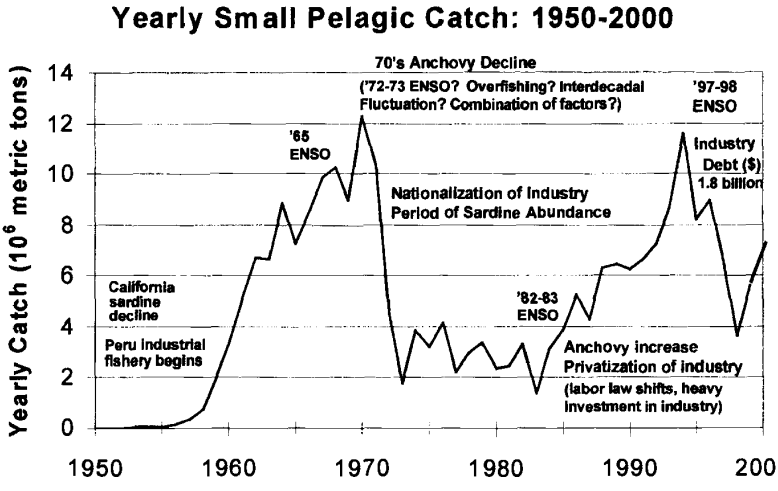


Figure 1 Peru annual small pelagic catch 1950–2000 (includes anchovy, sardine and mackerel). (Sources: 1950–1990: Csirke et al., "La ordenación y planificación pesquera y la reactivación del sector pesquero en el Perú" Rome: Jun 92. 1991–1996: Perú: desembarque de recursos marítimos, según especie 90-96 INEI: Jul 97. 1997–1999: *Statistical Reference Book*, No. 13: FEO Proceedings from 1999 FEO Annual Conference, Hong Kong, April 8–9, 1999. Paris: Fishmeal Exporters Organization.)

Research Programme (WCRP) under the auspices of the World Meteorological Organization (WMO), whose mandate resulted in national programs of various types around the world. Scientists were setting up an international experiment as part of CUEA, Coastal Upwelling Ecosystems Analysis, off the Peruvian coast when the 1972 El Niño event took place. Interest in forecasting El Niño was seen as socially relevant because forecasting the strength and biological productivity of coastal upwelling processes could be seen as more than an academic exercise (Glantz, 1979). It was suggested that knowledge gained about coastal upwelling and natural factors that inhibit it could be applied to address national economic development issues.

El Niño's impacts on Peruvian fisheries (especially the surface dwelling anchovy) became widely publicized during the 1972–1973 event, which has been blamed, along with overfishing and recruitment failure, for the collapse of that fishery. At that time about one-third of Peru's foreign exchange earnings were derived from the export of anchovy-based fishmeal. There are increasing claims, however, that the rapid decline in catch in 1973 began prior to the ENSO event, and that it was associated with natural fluctuations in abundance of small pelagic stocks. This is based on mounting evidence that anchovy (genus *Engraulins*) and sardine (genus *Sardinops*) populations fluctuate on multiyear or decadal scales as well as interannual timescales. Furthermore, there are indications of basin-wide synchrony in fluctuations of small pelagics (Bakun, 1996; Kawasaki et al., 1991; Lluch-Belda et al., 1992; Sharp and Csirke, 1983; Sharp and McLain, 1993).

The extraordinary 1982–1983 El Niño was the catalyst to expanding government and scientific interest in developing an El Niño forecast capability. In 1985, WCRP launched the multinational 10-year TOGA (Tropical Ocean Global Atmosphere) program that resulted in the recognition of ENSO as a key aspect in the interannual variability of the global climate system. The data collected from this project by researchers from as many as 40 countries aided the development of coupled ocean–atmosphere general circulation models. These models are intended to produce routine seasonal to interannual climate forecasts. Despite intensive worldwide coordination and effort invested in the physical understanding of ENSO phenomenon, only recently have researchers begun to address the socioeconomic factors. The rise in scientific understanding eventually translates into policy-making decisions and economic adjustments that have social consequences at all levels of society.

The ENSO event of 1997–1998 further heightened interest in the phenomenon and in the forecasting of it. Thanks to scientific concern and large-scale media coverage, numerous workshops and conferences, and the evolution of global communications technologies such as the Internet, societies around the globe became aware of El Niño and ENSO forecasts.

Peru, in particular, closely monitored ENSO information, and Peruvians generated their own forecasts of how the event would evolve. Peru has been a key member in the Comisión Permanente del Pacífico Sur (CPPS), which, in response to the 1972–1973 ENSO event, created a group called ERFEN (Estudios Regionales del Fenomeno El Niño). CPPS has committees in each of its member countries—Colombia, Ecuador, Peru, and Chile. Through this organization, national data

from cruises and observational stations are shared, allowing more comprehensive study of this regional phenomenon. In Peru, for example, several institutions, including the government oceanographic agency, the navy, the national meteorological service, the Peruvian Geophysical Institute, civil defense, and the private sector were involved in studies, conferences, and workshops attempting to enhance the understanding and preparation for the impacts of this event.

The Peruvian fishing sector, given its dramatic experiences with past events, was among the first groups to become concerned about the possibility of an extreme event in 1997–1998. Peru co-sponsored, along with the U.S. National Oceanic and Atmospheric Administration (NOAA), workshops bringing together international groups of scientists to discuss the matter, and to provide more detail about the event and its possible impacts. These workshops brought to light both the potential gains and difficulties in using forecasts for planning effective mitigative action in the fishing sector.

3 PERUVIAN FISHERIES SECTOR

In 1996, Peru was second to China in terms of the volume of fish landings and accounted for more than 10% of the world's catch. In 1997, this sector generated more than 2% of Peru's gross domestic product and consistently accounts for approximately 16% of all export products (second only to mining products). Bordered to the north by Ecuador and the south by Chile, the Peruvian coastline stretches 3100 km (01°01′48″ and 18°21′03″ south latitude), and has a continental shelf of 87,200 km^2. Its stark coastal desert is dotted with over 70 fishing ports ranging in population from a few hundred fishermen and their families to several hundred thousand persons. The fishing sector as a whole employs about 80,000 persons, out of a total population of over 24 million.

With the exception of ENSO years, the arid coastal climate is stable, as a result of the relatively cold coastal sea surface temperatures and high barometric pressure. There is evidence of reliance on living marine resources dating back to 7000 B.C., and continuing through the Moche, Chimbu, Nazca, and Paracas cultures, as well as during the Spanish colonial period up to the present. The focus, however, has changed from subsistence fishing for local consumption toward supplying a growing internal and international market. This makes fishermen not only reliant on local conditions and supply, which are impacted by the changing environment, but also vulnerable to fluctuations in global market prices and direct and indirect consumer preferences.

Peruvian coastal waters are home to over 107 commercial species (pelagic, demersal, and benthonic), of which 73 are fish, 11 crustacean, 16 mollusks, 2 echinoderms, and 5 algaes. Some of these species are classified as overexploited, while others are considered underutilized. Fluctuations in abundance of these species are a response to the environmental variability (often ENSO related) in combination with fish population dynamics, fishing pressure, habitat destruction, and pollution. Often, it is impossible to determine the precise reason for a population's variation

TABLE 1 Impacts of ENSO Warm Events on Common Marine Species

Marine Resource	ENSO (warm event)
Pelagics (e.g., primarily anchovy)	Start of event
	Schools concentrate near coast (easier to catch)
	Event strengthens
	Schools go deeper/migrate south
	Increased natural mortality
	Reduced reproduction/recruitment
Demersal (e.g., hake)	Go deeper/migrate south (harder to catch)
	Decreased natural mortality
Littoral (e.g., scallops, shrimp, octopus)	Population increases
Littoral (e.g., mussels, crabs)	Population decrease
Seabirds, marine mammals	Population decrease

Adapted from Ñiquen, M. et al., (1999). Efectos del Fenómeno El Niño 1997–1998 Sobre Los Principales Recursos Pelágicos en la Costa Peruana, In J. Tarazona and E. Castillo (Eds.), *Rev. peru. boil.* Vol. Extraordinario: 85–96.

because of the many factors in operation, in addition to the relative lack of knowledge of the life cycles of many species. What is clear is that some species are favored by the ENSO-related warm waters, while others are harmed. Table 1 summarizes some of the more consistent impacts on different species during an ENSO event.

Almost as varied as these marine organisms are those who harvest them. There are many types of fishers who use a range of equipment and techniques to gather the living marine resources. These fishers range from shore-based breathhold divers who use only a mask, fins, and speargun to shoot the large groupers (*Ephinephelus labriformis*) to the crews of the 800-ton industrial purse seine ships who use spotter planes and satellite images to search for schools of anchovy (*Engraulis ringens*) and sardine (*Sardinops sagax sagax*). Fishermen must register legally as artisanal (small-scale) or industrial, based on techniques, target species, and the size of their vessel. ENSO impacts these groups differently.

ENSO events shift the spatial availability and relative abundance of the species; a given event may benefit one member and harm another.

4 ARTISANAL SUBSECTOR

The artisanal subsector consists of more than 40,000 small-scale producers operating about 7000 vessels and can be characterized by the use of relatively rudimentary technology that has changed little over the last several decades. Artisanal fishermen use diving apparatus, nets, longlines, hook and line, and collect algaes and mollusks in the intertidal zones. Historically, they have occupied the lower socioeconomic strata of society, have been in general self-employed, and have had limited political influence because of poor organization as an economic subsector. Thus, their voice is manifested primarily through voting power.

Divers

Some divers simply hold their breath while hunting and gathering, while others use surface-supplied compressors to collect oysters and scallops, sea urchins, octopus and to spear a variety of finfish along the rocky shores in relatively shallow water (30 m or less). Their boats or rafts are usually rowed out to the fishing ground, powered by small outboards, by sail, or the fishermen simply dive from shore. The catch is then sold to middlemen back at the home port, where it is transported in refrigerated trucks throughout the country or shipped overseas to markets in the United States, Europe, and Asia.

Divers are generally adapted to the changing conditions of the water temperature brought on by ENSO. A moderate ENSO warm event, which warms the sea temperature, actually allows the divers to stay in the water longer without getting cold. As the water gets too warm, however, some species move into deeper waters, which lures the divers to follow them. Diving at deeper depths, combined with warm water that allows one to comfortably remain in the water for longer periods, can lead to an increase in incidence of decompression sickness, i.e., "the bends".

In extreme events such as 1982–1983, the water got so hot (more than 9°C above normal along the northern Peru coast) that many species of shellfish just died, while other species moved to depths outside the range of the divers. As diving equipment improves, and market demands for high-quality shellfish increase, divers will likely continue to push their depths. Unfortunately, training and emergency facilities are not on par with the increase in diving activity. This may be exacerbated, once awareness of an impending strong event occurs, as divers try to squeeze in as much time as possible in the water with the hope of "getting what you can" before the ENSO-related conditions deteriorate.

During warm events, however, some species such as octopus (*Octopus spp.*) and scallops (*Argopecten purpuratus*) grow at faster rates, which can permit increased harvesting at sustainable levels. Again, once the temperatures get too warm, however, these species can also perish. The temporary abundance of these commercially valuable species draws people from other occupations and areas of the country to begin diving for these marine resources. Most do not have proper training, which can result in increased diving accidents.

Net and Longline Fishermen

Net fishermen also fish for a range of finfish close to shore in small vessels (5 to 7 m), using gill nets that sometimes stretch over a kilometer in length. Once they have set their nets and are waiting for a couple hours to retrieve them, they will often fish with hook and line for bottom dwelling fish such as flounder (*Paralychthys adspersus*).

Longline fishermen tend to use larger vessels (10 to 15 m in length with inboard diesel motors), targeting shark, mahi-mahi (*Coriphaena hippurus*), and swordfish (*Xiphias gladius*). They often spend days at sea and go out as far as 80 miles offshore, depending on where the optimal water temperature and currents are found.

This group of fishermen is impacted by changes in the water temperature and depth of the thermocline, as tropical species move in closer to the Peruvian coast, making them more accessible to the fishermen. Mahi-mahi, for instance, feed on the eggs of flying fish (*Cypselurus heterurus*), and the flying fish are one of the first species to migrate with the warmer waters toward the coast, luring the mahi-mahi with them. During the early phases of the 1997 event, the abundance of mahi-mahi stretched down into northern Chilean waters, initially providing a steady source of income for the small-scale fishermen. However, as this ENSO began to increase in strength, an overabundance of these fish flooded the markets (both for internal consumption and for export), leading to a drop in prices. At one point, some fishermen stopped going to sea as the prices for their catch fell to very low levels ($1 per kilo); it was not worth their expenditures on fuel, ice, and materials.

The largest vessel of the artisanal fleet (approximately 30 gross registered metric tons) uses purse seine nets and is dedicated to the capture of anchovy, which is sold to the fishmeal plants. This group has a relative advantage during some ENSO events, as the anchovy move closer to shore in search of the nutrient-rich upwelled water, because the industrial fleet is not permitted to fish within 5 miles of the shore. This can lead to conflict and informal negotiations between the two sectors. With the total disappearance of anchovy during extreme ENSO events such as that of 1997–1998, some of these fishermen have modified their boats with permission and some minor subsidization from the government enabling them to trawl for langostinos, among other species that migrated down from Ecuador.

Many artisanal fishermen live in remote coastal rural villages and lack the infrastructure, such as ice machines and refrigeration systems, that enables them to store their products until the market situation improves. This makes them reliant on middlemen for the sale of their product, and often for ice, fuel, and other fishing supplies. This reliance is exacerbated during ENSO, as the temperature of the water, as well as of the air, can be several degrees Celsius above normal. Because their catches tend to spoil much faster due to the heat, they are more desperate to offload and sell their catch as soon as they reach shore.

The tendency for increased spoilage is problematic down the production chain. While it is difficult to blame directly on ENSO, there is a tendency for a sharp increase in gastrointestinal problems during the warm weather, again, due in part to a lack of refrigeration in many parts of the country, and the accelerated growth of bacteria because of adverse climatic conditions. An extreme example of the apparent linkage between ENSO and human health is the Pan-American cholera pandemic that began off the coast of northern Peru and spread across the continent in 1991. Facilitated by the 1992 ENSO, cholera destroyed the Peruvian market for artisanal fish and rapidly spread throughout South America. (Epstein, 1993). *Vibrio cholera*, which has been isolated from phyto- and zooplankton, is directly influenced by changes in water temperature and chemistry.

Another, perhaps equally important impact results from the increase in precipitation along the northern coast of Peru, where normally arid coastal communities can be inundated by the torrential rains. While fishermen may have an abundance of products to sell, the roads and bridges may be washed out by these rains and swollen

rivers, and, thus, there is no way to get their products to market. Again, as most fishing ports lack refrigeration, the products spoil due to the interruption in transportation.

Additionally, the sea level along the coast of Peru during an ENSO warm event may rise as much as 30 cm. The sea-level height is increased as a result of the eastward shift in ocean currents and the thermal expansion of the water that results from the increased ocean surface temperatures. This, in turn, can result in many more days with dangerous storm surges as waves from storms of "typical severity" have a stronger impact on the coast. ENSO-related storms, therefore, make it too risky for the small boats to navigate out of port.

5 INDUSTRIAL SUBSECTOR

The industrial subsector is characterized by a purse seine fleet of about 750 vessels with an average capacity of 225 metric tons, employing about 26,000 fishermen and plant workers. Most of these vessels target small pelagic species (primarily anchovy), of which more than 90% is processed into fishmeal, a flourlike substance high in protein that is then used throughout the world primarily as an animal feed supplement and in aquaculture farming. The majority of nonmanagement laborers are fishing fleet, fishmeal plant, and cannery workers. A significant number also work in associated industries such as net making and repair, engine repair, and shipping services. Most of the women employed in this industry work in the canneries. Throughout the period of commercial fishing, the industrial subsector has wielded strong political influence through lobbying activities and through explicit "places at the table" within the policy-making process.

ENSO warm events directly impact the anchovy stocks (there is a southern stock, which is shared with fishermen in northern Chile, and a larger north-central stock). During the onset of a warm event, the environmental conditions make the anchovy particularly vulnerable, as the pockets of cool, nutrient-rich water are reduced in area and are located near the coast. During the onset of an ENSO event, because the anchovy population becomes more concentrated near shore, they become easier to catch. The initial rise in catch is followed by a sharp decline, as the woresening conditions cause the fish to move deeper (out of the range of the nets) in search of food, and/or migrate southward in search of cooler waters. It is a combination of this spatial shift and fishing pressure that can rapidly reduce the short to midterm availability of stocks. In addition, during warm events, the reproduction and growth cycles of the anchovy are delayed and/or stunted, and in extreme cases, the fish may die because of poor environmental conditions. The impact of the reduction in anchovy population propagates through the food chain as, for example, and seabirds and marine mammals lose their food supply and are forced to dive deeper for food or to migrate southward. Widespread death of sea birds (e.g., guano birds) due to starvation is not uncommon during extreme ENSO events.

The short- and long-term impacts on the short-lived pelagic fish stocks are influenced not only by the current climate and fishing conditions but also by the state of the fish stock prior to the onset of an event, predation by other species, and other variables that are not directly ENSO related. For instance, if the stock is in good shape prior to an event, the odds increase that the stock will recover in a relatively short time period. The timing of the ENSO event may also influence the impacts. If the event peaks during the summer, as opposed to the winter, the absolute sea surface temperatures will be considerably higher, making a significant difference between slowed recruitment and reproduction versus actual survival of the organisms. However, in the case of extraordinary events like those of 1982–1983 and 1997–1998, the prior state of the stocks may be irrelevant and overwhelmed by the magnitude of change in environmental conditions.

The history of this fishery is one that illuminates the impact of ENSO events as well as the ensuing sociopolitical changes these climate shifts may spur. Prior to fishing, the anchovy stocks supported massive bird populations, whose excrement—guano—was the most effective natural fertilizer at the time. Fueled by the increased post–World War II demand for fishmeal and the collapse of the California sardine fishery (also blamed on overfishing combined with ENSO effects) (Radovich, 1981), the state-promoted Peruvian industrial fishing sector boom began in the mid-1950s and lasted until the early 1970s. In the context of weak regulations and technological advances, its catch increased to more than 12 million metric tons (primarily anchovy) by 1972. The ENSO-influenced anchovy collapse in 1973, coupled with political change in the country, led to a nationalization of the fishery, resulting in massive layoffs and a restructuring of the industry. During this epoch, the fishermen's labor union was much more influential and active, fighting against the nationalization of the sector in 1973 and then against denationalization in 1976 (Radovich, 1981).

In the 1970s and 1980s, the exploitation of sardines slowly replaced that of anchovies in the upwelling zone. The intense 1982–1983 ENSO event aided this regime shift, further reducing anchovy catches from an historic low of 118,168 metric tons in 1983 to 24,818 metric tons in 1984 tons. Sardine catches increased from 1,172,000 in 1983 to a peak of 3,398,000.

During the 1990s, however, anchovy reclaimed the ecological niche from sardine and returned to its primacy in commercial importance. Landings over the years up to the 1997–1998 ENSO event averaged more than 5,958,000 tons. The 1997–1998 event, however, led to a decline in catch from over 8 million metric tons in 1996 to over 6.6 million metric tons in 1997 and less than 4 million metric tons in 1998, and almost 6 million metric tons in 1999.

During ENSO warm events, some of the fish migrate southward to relatively cooler waters. Fishmeal plants in the southern section of the country increase catch (and profits), while the plants to the north face difficult times. Some firms have adapted by building plants in the north, central, and southern parts of the coast to take advantage of the spatial fluctuations in resources. Others have decided only to operate fleets, which they can then move up and down the coast as the movement of the resource dictates. Other firms own both plants and fleets, which allows them to

better hedge the supply in the face of uncertainty. A larger scale economic response to the high variability of the fish stocks by the largest fishing firms is diversification into canned fish products, agriculture, mining, and other industries.

The few years preceding the 1997–1998 ENSO event were characterized by increasing catches combined with high fishmeal prices that spurred large investments in new boats and plants (as occurred in the years prior to the 1972–1973 event), financed by loans from private banks. Overcapitalization coincided with the onset of the 1997–1998 event and despite a relatively rapid recovery of the anchovy stocks, low fishmeal prices have greatly reduced the financial recovery of the industry. In terms of labor, unlike in the early 1970s, union power has at present virtually disappeared from the fishing sector, having been replaced by the large fraction of short-term contract (as opposed to permanent) workers. The government, following an extreme neoliberal philosophy has not been willing to subsidize the industry.

Many of Peru's banks have major investments in the industrial fishing sector (at the end of 1996, the industry as a whole had borrowed over US$270 million and by the start of 1999 the industry was alleged to owe the banks approximately US$1.5 billion). As ENSO events begin to evolve, banks must make decisions on whether to make new loans. Once an event is underway, they will often refinance loans on terms based upon expectations of the upcoming quarter's catch. The banks respond to a range of information, including rumors and media coverage. Some banks hired scientists from the government agencies as consultants and co-sponsored ENSO forums to evaluate the event and its impact on the fishing sector. Despite the early awareness of the 1997–1998 event, the recent history of overinvestment in vessels and plants led to major defaults on loans by virtually all the fishing firms. This exacerbated the countries' economic crisis by adding to the unpaid debts of the agriculture and mining industry, and several large banks were forced to close down or merge with other banks to survive. By the end of 1999, the catch had returned to fairly good levels. However, the fishing firms' economic situation could not be remedied and the industry remained in a crisis. At the time of the writing of this chapter, the government is discussing the need for a major restructuring of the industry, including reducing the fleet size and number of plants, and implementing additional regulatory mechanisms such as individual transferable quotas.

Regulations are made by the Ministry of Fisheries. In theory, its decisions are informed by the recommendations of the board of directors of the governmental scientific organization in charge of fisheries and oceanographic studies. Regulatory mechanisms include species as well as minimum size restrictions, closed seasons (*vedas*), spatial as well as gear restrictions, and statistical reporting. These are not consistently enforced as regulators face difficult decisions during the various stages of ENSO. During the 1997–1998 event, the Ministry of Fisheries implemented several short additional fishing bans to protect some resources (though they were heavily contested by the industry) and actually passed a special decree granting licenses to allow fishermen to temporarily fish jack mackerel (*Trachurus murphi*) and chub mackerel (*Scomber japonicus*) with sardine nets (smaller size mesh than normally allowed) during this event. Thus, there is a constant trade-off between trying to conserve the species, preserve jobs, and appease the politically powerful

industrial interests. It is in this context that ENSO forecasts can become a key factor in public-sector policy formation.

6 POLICY IMPLICATIONS OF CLIMATE INFORMATION

In the previous sections, we identified some of the positive and negative socio-economic effects and adaptations by members of the Peruvian fishing sector to ENSO events. At the artisanal level, this may include gear switching or migration to take advantage of changing marine resources. At the industrial level, many firms have adapted by diversifying, including moving into canning as well as fishmeal, but also into other industries. Some industrial fishermen and plant workers have second jobs, or small shops, often run out of their homes by family members. This helps to carry them through the leaner fishing seasons, as there are no laws that guarantee a minimum wage or labor security during poor fishing periods. Banks respond by altering their loan policy (usually in unfavourable ways) and by refinancing existing loans. Scientific institutions adapt by increasing their monitoring efforts, as well as using the event itself to lobby for more funding from the central government. Regulatory agencies may increase vigilance, change gear restrictions, or in some cases alter quotas.

Secondary impacts on the sector include the increased spoilage and health effects induced by the high air temperatures and transportation problems caused by intense rains. On a macroeconomic level, ENSO events favourably affect the growing, marketing, and sale of soy bean products in other parts of the world. Soymeal is the main competitor with fishmeal as an animal feed supplement. Buyers chose between these two meals, based on their relative prices.

Given the range of impacts and adaptations that are made in response to climate variability, combined with the substantial improvement in understanding and predicting the ENSO phenomenon, one can imagine more efficient political and private-sector decision-making possibilities. Observations from the 1997–1998 ENSO highlight some of the challenges in policy formation based on climate forecast and information. For a thorough treatment of this point using the case of fisheries and food security, see Broad et al. (2002) and Broad and Agrawala (2000). These difficulties stem from two primary sources: (1) scientific uncertainty and (2) societal constraints on the use of climate information.

Scientific Uncertainty

One of the primary constraints is that climate forecasters have relatively poor skill in predicting sea surface temperature (SST) and thermocline depth near coastal areas due to the steep gradients in oceanographic as well as bathymetric properties, compared to the open ocean areas. Thus, it may be difficult for regulators to plan their closed seasons and scientific cruises with much precision because of this temporal and spatial uncertainty. One of the factors preventing the more accurate prediction of ENSO impacts on the coastal ecosystem of Peru is the lack of regional

scale (e.g., coastal) observations. Enhanced observations, combined with the targeted training and further model development, would aid in the monitoring and prediction of changes in coastal environmental conditions. Further, even if there were perfect climate forecasts (an impossibility), there remains the difficult step of incorporating such information into fisheries models (Glantz, 1979).

In addition to climate there are numerous variables that impact fish populations, including fishing pressure. Year-to-year and multitimescale variability in stock abundance during non-ENSO years is still not well understood (Bakun, 1996; Sharp and Csirke, 1983; Kawasaki et al., 1991; Lluch-Belda et al., 1992). One example of how this uncertainty is played out in real decision making is highlighted by the two options that were being debated during the 1997–1998 ENSO event. On the one hand, at the start of the event, conservationists called for a halt to all fishing in order to preserve as many anchovy as possible in the hope of preventing a future total collapse of the fish population. A complete ban on fishing, of course, has massive social implications, including widespread unemployment, loss of export dollars, and ensuing political unrest. This regulatory move may be rationalized by claiming that it is an attempt to ensure resources for current as well as for future generations. On the other hand, as it became clear that an extremely warm event was underway, some argued, not without reason, that fishing regulations should be removed and the fishermen should be allowed to catch what they can before all the fish disappear for "natural" reasons. Again, an arguably "rational" position, albeit self-serving, based on past experience during extreme ENSO events.

To decide the best option between these two possibilities (or a combination thereof), there is a need for more reliable forecasts as well as improved fisheries population dynamics models. Only recently have meteorologists, biological oceanographers, and fisheries biologists begun to join forces on these problems. During the 1997–1998 event, the Peruvian authorities followed a more politically viable middle ground, by exercising what can be considered "adaptive management" practices, i.e., adjusting their standard regulatory measures based on increased observations and sampling, allowing them to set their closed fishing periods with greater flexibility in response to a changing environment and political pressure.

Societal Constraints

Members of the fishing sector have different, often conflicting, goals. For instance, most industry members want to maximize their short-term profits; regulators want to conserve the resource for future generations, but also want to keep their jobs (as they recognize the power of the industry to get them fired); and banks want positive returns on their investment. These conflicting goals lead to the "political" use of the scientific information such as climate forecasts. For example, during the 1997–1998 ENSO event, those industrial fishing firms that had many outstanding loans tried to coerce some scientists to claim that it was going to be a strong ENSO with a very negative impact on the fisheries. Their logic was that the banks would then be willing to refinance under more favorable conditions or temporarily suspend interest payments. In contrast, firms waiting for loans pressured scientists to downplay the

potential severity of the event in order not to scare the bankers out of making the loans for fear of a collapse and several years of low catches. In addition, two of the largest firms had begun issuing bonds immediately prior to the event and were afraid of scaring off potential investors with doomsday predictions of a fish stock collapse.

Attempts to sway economic decisions based on expectations of the evolution of the ENSO event were played out in the media, which also capitalized on the sensational aspect of a looming "disaster." The media also served as the venue for competing forecasts issued by different Peruvian scientific institutions, each one of which was vying to be the voice of authority on the event. A successful forecast could help them to get funds from the state to further research and monitor the event. With public servant salaries relatively low, individual scientists were tempted by industry's incentives—often negotiated under the table—to interpret uncertain information for the benefit of industry objectives.

Another constraint on the management of the Peruvian fisheries in light of the 1997–1998 ENSO event results from geopolitical border politics. Peru shares the southern anchovy stock with Chile, and during ENSO warm events, the stock tends to migrate further southward into northern Chilean waters. As the Peruvian quota was reached in the south, there was extreme pressure put on the regulators *not* to halt fishing in the south, the argument being that Peru should not stop fishing, since the fish were just going to go to their Chilean competitors. After much heated debate, and being accused of antinationalist tendencies, the Peruvian authorities did ban fishing in the south for a period of time, although it was only after increasing the original quota.

When considering the potential impact of forecasts, an additional factor that arises is that some groups may be more susceptible to negative impacts of climate forecasts and *mis*forecasts than others. For example, a small-scale fisherman who normally fishes close to shore and has little personal savings may receive a forecast of an ENSO event and decide to sell his gillnet and buy a longline net in anticipation of the arrival of tropical species. If the event does not occur with the anticipated intensity, however, he is left with useless gear that no one will want to buy. An example from the industrial sector may be that, if owners have prior information that fishing may be poor in the upcoming months, they may fire plant workers in advance in order to reduce their potential losses. For a discussion of who may benefit from climate forecasts at the expense of others, see Pfaff et al. (1999).

These scenarios suggest that the way climate information is disseminated plays a critical role in how society makes use of it. A forecast provider should be aware of issues of equity when making a forecast publicly accessible. Merely putting information on the Internet, for instance, may allow access to only a limited few in Peru, such as the industrial owners and managers. In contrast, many of the rural fishing villages receive information via short-wave radio, which presents a different dissemination challenge for forecast providers. Groups also have differing capacities to understand probabilistic forecasts, thus necessitating that varied types of training and education accompany the distribution of these forecasts.

We have entered a new phase in meteorology, where operational climate forecasts are being generated, and much effort is being put into disseminating this information

in the United States and abroad. If any country stands to gain from this information, it is Peru, a nation directly impacted by ENSO-driven interannual climate variability. Maximizing the societal value of this information, however, necessitates communication and cooperation by those generating the predictions, distributing the forecasts, and the end users of this information. Only this approach can assure the equitable distribution of information in the proper form, with the proper training, to various decision makers at different socioeconomic strata of society.

REFERENCES

Arntz, W., A. Landa, and J. Tarazona, *Boletín: "El Niño": Su Impacto en la Fauna Marina* (Volumen Extraordinario), IMARPE, Callao, 1985.

Bakun, A., *Patterns in the Ocean: Ocean Processes and Marine Population Dynamics*, California Sea Grant College System, 1996.

Barber, R. T., and F. P. Chavez, Biological consequences of El Niño, *Science*, *222*, 1203–1210, 1983.

Bjerknes, J. A possible response of the atmospheric Hadley Circulation to equatorial anomalies of ocean temperature, *Tellus*, *8*, 820–829, 1966.

Bjerknes, J., Atmospheric teleconnections from the equatorial, *Monthly Weather Rev.*, *97*, 164–172, 1969.

Broad K., and S. Agrawala, The Ethiopia food crisis: Uses and limits of climate forecasts, *Science*, *289* (8), 1693–1694, 2000.

Broad, K., et al., Effective and equitable dissemination of seasonal-to-interannual climate forecasts: Policy implications from the Peruvian fishery during El Niño 1997–98, *Climatic Change*, 1–24, 2002.

Carr, M.-E., and K. Broad, Satellites, society, and the Peruvian fisheries during the 1997-98 El Nino, in D. Halpern (Ed.), *Satellites, Oceanography and Society*, Elsevier Science B.V., Amsterdam, 2000, pp. 171–191.

Chavez, F. P. et al., Biological-physical coupling in the central equatorial Pacific during the onset of the 1997-98 El Niño, *Geophys. Res. Lett.*, *25*, 3543–3546, 1998.

Epstein, P. R., Algal blooms in the spread and persistence of cholera, *BioSystems*, *31*, 209–221, 1993.

Glantz, M. H., Science, politics, and economics of the Peruvian anchoveta fishery, *Marine Policy*, 201–210, 1979.

Kawasaki, T., S. Tanaka, Y. Toba, and Taiguchi, *Long-Term Variability of Pelagic Fish Populations and Their Environment: Proceedings of the International Symposium*, Pergamon, 1991.

Lluch-Belda, D., et al., Sardine and anchovy regime fluctuations of abundance in four regions of the world oceans: A workshop report, *Fish. Oceanogr. 1*, 339–347.

McPhaden, M. J., Genesis and evolution of the 1997–98 El Niño, *Science*, *283*, 950–954, 1999.

Pfaff et al., Who benefits from climate forecasts? *Nature*, *397* (25), 645–646, 1999.

Philander, S. G., *El Niño, La Niña, and the Southern Oscillation*, Academic, New York, 1990.

Quinn, W. H., V. T. Neal, et al., El Niño occurrences over the past four and a half centuries, *J. Geophys. Res.*, *92* (14), 1449–1461, 1997.

Radovich, J., The collapse of the California sardine fishery: What have we learned? in M. H. Glantz and J. D. Thompson (Eds.), *Resource Management and Environmental Uncertainty: Lessons from Coastal Upwelling Fisheries*, Vol. 11, Wiley, New York, 1981, pp. 107–137.

Rodbell, D. T., et al., An ~ 15,000-year record of El Niño–driven alluviation in Southwest Ecuador, *Science*, *283*, 516–520, 1999.

Sharp, G. D., and J. Csirke, *Proceedings of the Expert Consultation to Examine Changes in Abundance and Species Composition of Neritic Fish Resources*, Food and Agriculture Administration, San Jose, Costa Rica, 1983.

Sharp, G. D., and D. R. McLain, Fisheries, El Niño–Southern Oscillation, and upper-ocean temperature records: An Eastern Pacific example, *Oceanography*, *6*, 13–22, 1993.

CHAPTER 45

DROUGHT IN SOUTH AFRICA

COLEEN VOGEL

1 INTRODUCTION

Drought is a recurrent phenomenon in Africa with occurrences noted over several centuries, varying in spatial extent and severity (Glantz, 2001; Nicholson, 1978; 1989). The impacts of drought are diverse occurring at several scales including at household, national, and regional levels. When droughts are coupled with poor mitigation strategies and lack of preparedness, accentuated periods of hardship can occur particularly for poor, vulnerable, urban, and rural areas (e.g., Glantz and Katz, 1985; Vogel, 1995; Davies, 1996; Downing et al., 1996). Dry spells, heavy rains, disruption to commercial farming, depletion of grain reserves and increases in staple food prices are some of the factors contributing to food insecurity in Southern Africa during 2002 (World Food Programme, 2002). Drought is therefore a multifaceted phenomenon, the consequences of which are the result of complex human and biophysical interrelationships. In this chapter these interactive dimensions of drought are traced for South Africa.

2 BIOPHYSICAL DIMENSIONS OF DROUGHT IN SOUTH AFRICA

Several factors interact to produce the climate and weather of South Africa (for good overviews see Tyson, 1986; Mason and Jury, 1997; Lindesay, 1998; Tyson and Preston-Whyte, 2000). The physical location of the country in the subtropics shapes climate and weather (Fig. 1). The local orography and landscape features of the country (most of the country is situated on a plateau approximately 1500 m above mean sea level) and sea surface temperatures (SSTs) of the cooler Atlantic and warmer Indian Ocean also modulate and influence local weather and climate.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

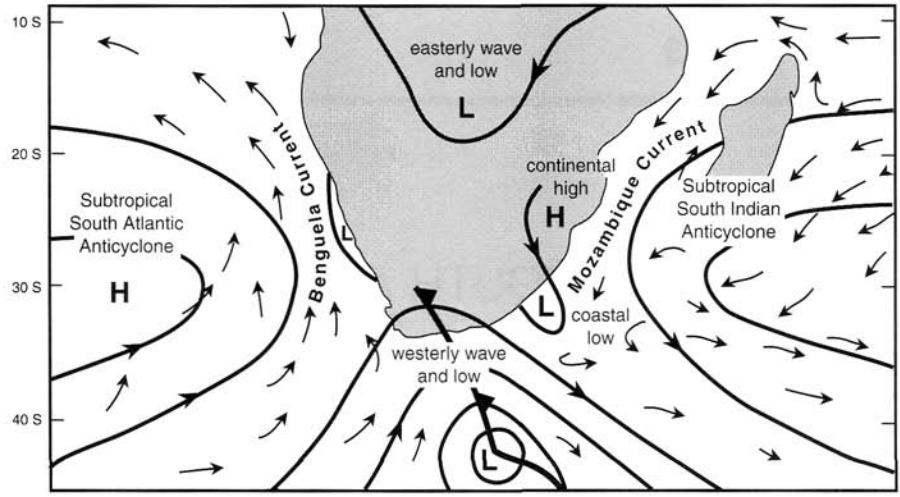


Figure 1 Some important features and controls of the atmospheric circulation over southern Africa (modified after Tyson and Preston-Whyte, 2000, with kind permission, Oxford University Press, Cape Town).

The subtropical high-pressure belt (e.g., the South Atlantic High and the South Indian Anticyclone) is usually associated with dry, stable atmospheric conditions that can prevail over much of the country for several days of the year (Tyson and Preston-Whyte, 2000). Tropical easterlies also affect southern Africa throughout the year (Tyson and Preston-Whyte, 2000). Sources of atmospheric moisture from the warmer tropical regions interact with colder drier air from the southwest to provide the setting for weather producing systems. The interaction between warm, moist easterly air and drier, colder westerly air produce convergence zones, cloud bands, and rainy spells for days across the country. The variation and location of these cloud-band convergence zones over the southern Africa–Indian sector in summer has important implications for rainfall over the country (Tyson and Preston-Whyte, 2000; Harrison, 1984a).

Wider-scale interactions between oceans (sea surface temperatures) and the atmosphere also influence the variability of rainfall (Jury, 1995; Jury et al., 1996; Mason and Jury, 1997), both locally and on a larger scale. Changes in the sea surface temperatures that are linked to atmospheric pressure fields and circulation have been shown to influence local weather and climate (Tyson and Preston-Whyte, 2000). Warmer and colder phases of water in the Pacific Ocean, for example, may influence the climate of the country by enhancing or suppressing local circulation systems. As will be shown below, these changes in circulation can become pronounced (e.g., during an El Niño or La Niña) and can produce periods of extremes in temperature and rainfall over the country (Lindesay, 1998).

The combination of the biophysical features outlined above configures the country into arid and semi-arid areas in the west of the country, becoming wetter as one

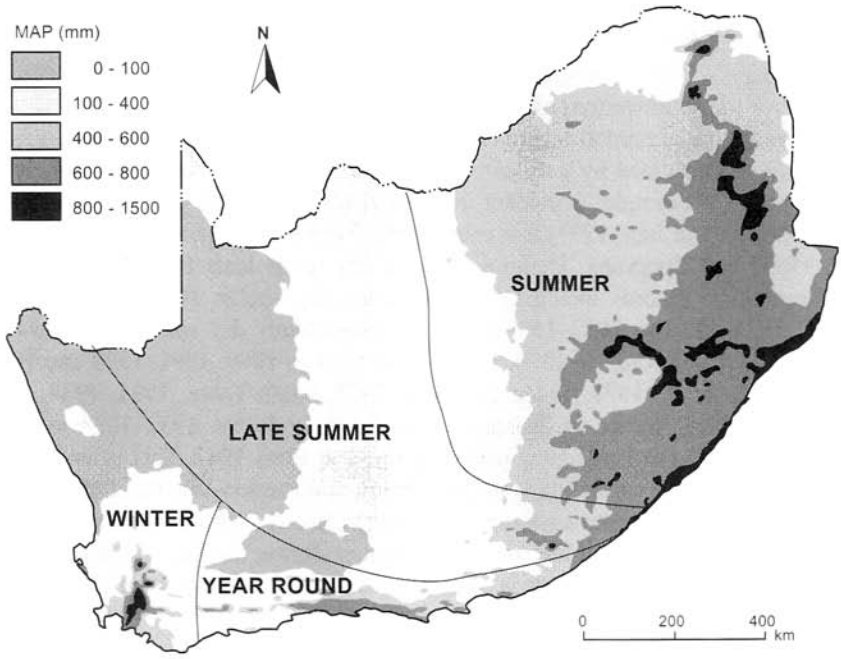


Figure 2 Spatial distribution of mean annual rainfall in South Africa (after Fox and Rowntree, 2000, with kind permission, Oxford University Press, Cape Town).

progresses eastward (Fig. 2). The surface aridity of much of the country is a function of the variability of the rainfall* as well as the relationship between evaporation and precipitation (Lindesay, 1998; Tyson, 1986). Evaporation usually exceeds precipitation over almost all of southern Africa with the deficit being greatest in the west, particularly in western South Africa and neighboring Namibia and Botswana. As a consequence of the topography and rainfall distribution, the natural availability of water is very unevenly distributed across the country with more than 60% of the flow arising from only 20% of the land area (Department of Water Affairs and Forestry, 1997; Schulze et al., 2001). Large bulk water transfer schemes have been established to ensure water supply from these wetter to drier areas (Abrams, 1997).

3 RAINFALL VARIABILITY

Rainfall in southern Africa varies temporally with annual, seasonal, and daily variations in the amount of rainfall received. Southern Africa, for example, experiences a high level of intraannual and interannual rainfall variability (Tyson, 1986). Rainfall occurs in the summer months across the central and northeastern parts of the country

* Precipitation usually includes the deposition of water in solid or liquid form including rain, dew, snow, hail (Goudie, 1994). In this chapter rainfall will be the main focus of the discussion.

with a winter rainfall area predominating to the southwestern part of the country. Summer rainfall is usually characterized by frequent thunderstorms over much of the central parts of the country, and winter precipitation is usually the result of frontal rains over the southwestern areas.

Assessments of rainfall records for South Africa indicate that interannual rainfall has been characterized by a series of wetter and drier years. An approximately 18-year cycle, for example, is evident over much of the summer rainfall area (Tyson, 1986; Mason and Jury, 1997) that extends into Zimbabwe (Torrance, 1972; Ngara et al., 1983) and Botswana. Definite wet and dry spells have been identified from meteorological records including the following dry spells: 1905–1906 to 1915–1916; 1925–1926 to 1932–1933 (the most consistently dry spell), 1944–1945 to 1952–1953; 1962–1963 to 1970–1971; 1980–1981 to 1990; 1991–1992 and 1994–1995 (Tyson et al., 1975; Tyson and Dyer, 1978, 1980; Tyson, 1981, 1984, 1986; Lindesay, 1998). Wetter spells occurred in the early 1920s, 1933–1934 to 1943–1944, 1953–1954 to 1961–1962 and 1971–1972 to 1980–1981. This latter wet spell was also the most persistently wet spell (Tyson and Preston-Whyte, 2000). Though drier spells tend to exhibit a greater spatial homogeneity than wet spells, the impression should not be gained that rainfall anomalies during the wet and dry spells are spatially uniform. Rather there is marked interannual and spatial variability in the distribution of wetter and drier conditions (Tyson, 1986; Lindesay, 1998). Available evidence suggests this pattern of wet and dry years has generally persisted since at least 1840 (Vogel, 1989).

4 CAUSES OF DROUGHTS IN SOUTH AFRICA

Having described the pattern of rainfall-producing systems in the country, attention now focuses on some of the causes of periods of insufficient rainfall and droughts. The forcing mechanisms for droughts are well documented for the southern African region (see, e.g., Tyson, 1986; Mason and Jury, 1997; Lindesay, 1998; Tyson and Preston-Whyte, 2000). Much of the recent research in the country has focused on the prolonged droughts of the last two decades (namely the 1980s and 1990s) Tyson and Dyer, 1978; Dent et al., 1987; Jury and Levey, 1993; Alexander, 1995; Mason and Jury, 1997; Lindesay, 1998; Landman and Mason, 1999).

Rainfall variability for the country has been associated with atmospheric circulation configurations and interchanges in easterly and westerly circulations, the interactions between tropical and temperature systems, and the variation in pressure patterns over Marion and Gough Island (summarized in Tyson, 1986; Lindsay, 1998). Prolonged heat waves and droughts are linked, in most cases to the predominance of prevailing anticyclonic circulation over the country. Longer periods, such as extended wet spells, for example, are usually caused by an invigoration of tropically induced circulation disturbances forced by tropical easterlies whereas the extended dry spells usually occur with an expansion and increase of westerly disturbances. In the latter case, summers become drier in the summer rainfall region (Tyson and Lindesay, 1992). Aspects of this circulation patterning during wet and

dry spells have been shown over long time scales, possibly during late Interglacial and mid-Holocene warming phases (Partridge, 1997).

Much of the work on seeking causal mechanisms for rainfall variability has focused on the relationship between rainfall and the circulation parameters within the region, such as changes in the intensity and positioning of pressure systems (Tyson, 1981, 1984; Miron and Lindesay, 1983; Taljaard, 1986, 1989). Adjustment in factors such as cloud bands (important conduits of energy and momentum into the subcontinent) and ocean temperature (Walker and Lindesay, 1989; Jury, 1995) have been shown to also induce rainfall-related changes over the subcontinent and more specifically over South Africa (Harrison, 1984a, b, c; Lindesay, 1988; Mason, 1990).

More recent work has also increasingly focused on the global-scale forcing mechanisms of rainfall and linkages to atmospheric circulation changes across the Southern Hemisphere. The *Southern Oscillation* (the climatic oscillation between warm and cold periods in the tropical Pacific) and *sea surface temperature* (SST) in the South Atlantic and South Indian Oceans, have become key foci of climate research in recent years (e.g., Tyson, 1986; Lindesay, 1988, 1998; Harrison, 1986; Jury et al., 1996; Mason and Jury, 1997; Landman and Mason, 1999). Aspects of these, relating specifically to drought, are highlighted below.

The Southern Oscillation is a major forcing mode of the interannual circulation variations over southern Africa (Lindesay, 1988; Allan et al., 1996). One mechanism by which the Southern Oscillation signal is transmitted to southern Africa is via the Indian Ocean Walker Circulation. High- and low-phase years of this circulation (known as El Niño Southern Oscillation, ENSO) have been shown, furthermore, to influence rainfall over southern Africa (e.g., Lindesay, 1988, 1998). Briefly, low-phase years of the El Niño Southern Oscillation are accompanied by reductions in heat release and convection over tropical southern Africa caused by adjustments in the Walker Circulation and a retarded number of tropical-temperate troughs across South Africa. Fewer cloud bands occur and rainfall is diminished (Harrison, 1986). The reverse essentially occurs for high-phase or wetter years (Fig. 3). This general association between the Southern Oscillation phase changes and South African rainfall has been shown to exist in both present and preinstrumental rainfall records in South Africa from at least 1820 (Lindesay and Vogel, 1990).

An integral part of the Southern Oscillation is the role that sea surface temperatures play in modulating the occurrence of low- and high-phase changes and associated atmospheric circulation interactions. Although not explicitly direct, relationships between sea surface temperature anomalies, atmospheric circulation, and rainfall have been shown to exist in several instances (Glantz et al., 1987; Ogallo et al., 1988; Nicholson and Entekhabi, 1986; Mason, 1990; Walker, 1990) including relationships between above-average sea surface temperatures in the Benguela area and dry years such as 1982–1983 (Walker et al., 1984; Philander, 1986).

ENSO warm events have been associated with drought, resulting in a variety of impacts over much of southern Africa (e.g., Ogallo, 1987; Enfield, 1989; Cane et al., 1994). The 1982–1983 ENSO event, for example, served to exacerbate the prevailing dry conditions in much of the subcontinent (Bhalotra, 1985; Dent et al., 1987; Taljaard, 1989). The rainfall-producing systems of the subcontinent were displaced

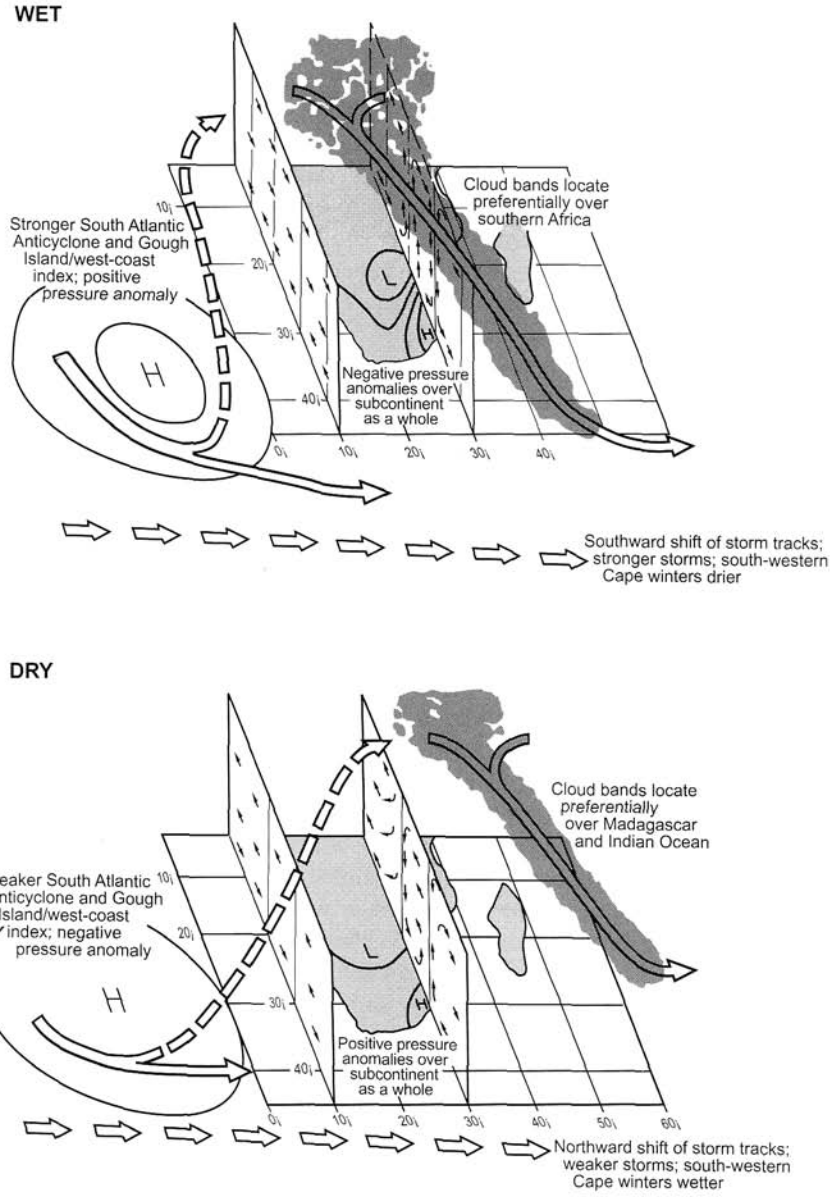


Figure 3 Model of the circulation changes over southern Africa during wet and dry spells (after Tyson and Preston-Whyte, 2000, with permission, Oxford University Press, Cape Town).

eastward during this time (Fig. 3) (Harrison, 1983; Tyson, 1986; Lindesay, 1988; Muller and Tyson, 1988).

The recent ENSO event of 1997–1998 was a very powerful one with anomalously high SSTs and, thus, indications for a possible drought (of similar magnitude to that of 1982–1983); interventions were planned, both locally and elsewhere (see Thomson et al., 1998; Mason et al., 1999; and NOAA-OGP, 1999). The El Niño impact was reduced in some parts of southern Africa and appears to have been modulated by temperatures in the Indian and Arabian Sea (Landman and Mason, 1999). Despite the strong linkage between ENSO and rainfall in parts of southern Africa, it must be remembered that not all droughts are associated with ENSO. There thus remain areas of uncertainty and predictive skill in seasonal forecasting for the country because of the complex interactions between the oceans, atmosphere, and land (Mason et al., 1994, 1996; Landman and Mason, 1999).

In summary, current research on the causes of droughts in South Africa indicates that rainfall over the country is influenced by a number of interactive mechanisms, details of which are still being examined (Mason and Jury, 1997; Landman and Mason, 1999; Mason and Tyson, 2000). The interaction between tropically and extratropically sourced weather systems, when combined with upper-air tropospheric dynamics and convectively unstable air masses, usually results in high possibilities of rainfall. Extreme drought events in southern Africa are the result of a number of atmospheric circulation interactions (Tyson, 1986; Lindesay, 1998) with some droughts and periods of reduced rainfall, for example, connected with ENSO events.

5 CLASSIFYING DROUGHTS

With this background, the focus narrows to consider the character of droughts and drought impacts in South Africa in more detail. Defining what is meant by drought is not an easy task. Droughts can be classified as meteorological, agricultural, hydrological, or sociological (Wilhite and Glantz, 1985; Dent et al., 1987; Erasmus, 1987; Bruwer, 1989, 1990). Drought, moreover, is also a relative rather than an absolute condition. Droughts can be described as being either an agricultural drought (a condition where soil moisture is depleted such that yields are considerably reduced) or a hydrological drought (actual water supply being less than the minimum required for normal operations) (Wilhite and Glantz, 1985).

In South Africa, drought is broadly defined as occurring when 75% or less of normal precipitation is received (Laing, 1992), being classed as severe if it extends over two consecutive seasons. Other indices such as the Palmer Drought Severity Index (PDSI), include inputs of soil moisture, runoff, evaporation, and temperature (Zucchini and Adamson, 1984; Erasmus, 1987; Wilhite, 1987; Bruwer, 1990). Classification of a dry spell and/or drought depends on the definition, criteria, and statistical methods used.

The timing of rainfall is also critical in determining the progression and consequences of a drought situation. Although most of the country, for example, experienced good rainfalls in 1987–1988 and 1988–1989, the individual rain events were

short, torrential downbursts. As these incidences show, timing and adequate rainfall, balanced against evaporation and infiltration, determine the amount of rainfall available for surface and groundwater usage. Opportune rainfall is also essential for crop growth, and several improvements in determining drought severity and crop response have been developed including the Agricultural Catchments Research Unit (ACRU) model (Schulze, 1984, 2000; Schulze et al., 2001) and the Crop Environment Resource Synthesis (CERES-MAIZE) model (Du Pisani, 1987).

Using such methods, together with other descriptive criteria including rainfall over three seasons, grassland condition, availability of water for stock, stock condition/deaths, availability and volume of fodder purchased (Bruwer, 1989, 1990), one is able to divide the country into areas that are more or less drought prone. Records for a 30-year period, for example (1936 to 1986), show that 27% of the country has been declared a disaster drought area for more than 50% of the time (Bruwer, 1989, 1990).

Periods of drought, usually spanning at least 2 or more seasons, occurred in the late 1920s and early 1930s, much of the 1960s and 1980s, and more recently in the early and mid-1990s. These drought periods have brought with them and have compounded a variety of problems (including access to water in rural poor areas and water availability and use in agricultural and industrial sectors). Droughts have also highlighted several prevailing factors that predispose certain areas and groups to heightened drought risk (e.g. poor water infrastructure, land degradation, globalization) (Vogel, 1994; Scoones et al., 1996; Benson and Clay, 2000; Dilley, 2000).

6 IMPACTS OF DROUGHTS IN SOUTH AFRICA

Drought impacts, however, are not the result only of insufficient rainfall or searing temperature. In most cases, drought impacts are the outcome of the interaction of a number of social and other human factors that can heighten the "vulnerability" of communities and various exposure units (e.g., vegetation) and reduce "resilience" of society and ecosystems to the natural hazard (Dilley, 2000; Vogel et al., 2000). As a result of these components of drought, a number of impacts are recorded.

The scale of these impacts also varies and can be tracked at various levels (e.g., regional, national, community, and household) of agricultural production. For example, production declined as a result of the 1980s and 1990s droughts in southern Africa. Harvest failures of between 30 and 80% below-normal across the Southern African Development Community (SADC) region were recorded. Cereal production in the SADC countries dropped to less than 50% of the annual requirement in 1992, and the cost of imported food to the region rose to approximately $4 billion (Hulme, 1996). Drought related to the 1982–1983 El Niño cost nearly US$1 billion in direct damages with an estimated US$350 million spent on famine relief (1983 prices) in southern Africa (International Federation of Red Cross and Red Crescent Societies, 1999). The economic loss to Africa's agricultural sector in the early 1990s drought was estimated at US$7 billion (1992 prices)–an estimated 20 times the value of 1993 World Bank loans to sub-Saharan agriculture (International Federation of Red Cross

and Red Crescent Societies, 1999). More recently, the combined influence of drought, floods, economic instability, and HIV/AIDS threatened the food security of millions in southern Africa (WFP, 2002).

At a national level, drought ripples across sectors and impacts on a range of activities. On a national scale, drought in South Africa results in a reduction in the yield of the maize crop with yields falling to below 1 ton per hectare (Fig. 4). The outward effects triggered by drought range from its impact on agriculture's contribution to the gross domestic product (GDP) to a host of other impacts such as food supply, employment opportunities, and a number of forward and backward linkages associated with the agricultural sector (Ballard, 1986; Van Zyl et al., 1987; Van Zyl and Nel, 1988). The declining agricultural yields associated with the droughts of the 1990s, for example, negatively affected GDP growth by between 0.5 and 2% (Mather and Adelzadeh, 1997).

The human consequences and across sectors and groups, are difficult to quantify accurately. For the agricultural sector the occurrence of drought, together with changes in agricultural policy, provision of farmer loans and other economic factors, can combine to heighten the impacts on the sector. During the recent severe drought of the early 1990s, for example, it was estimated that 50,000 jobs would be lost in the agricultural sector (with a further 20,000 in related sectors) and about 250,000 in total (families included) would be affected (AFRA, 1992; Adams, 1993; Van Zyl, 1993). Crop failures occurred for both commercial and smaller-scale farmers (Adams, 1993) and water levels in several of the major dams were less than two-thirds their normal capacity (Fig. 5). Faulty and poorly maintained water infrastructure further aggravated the precarious water situation.

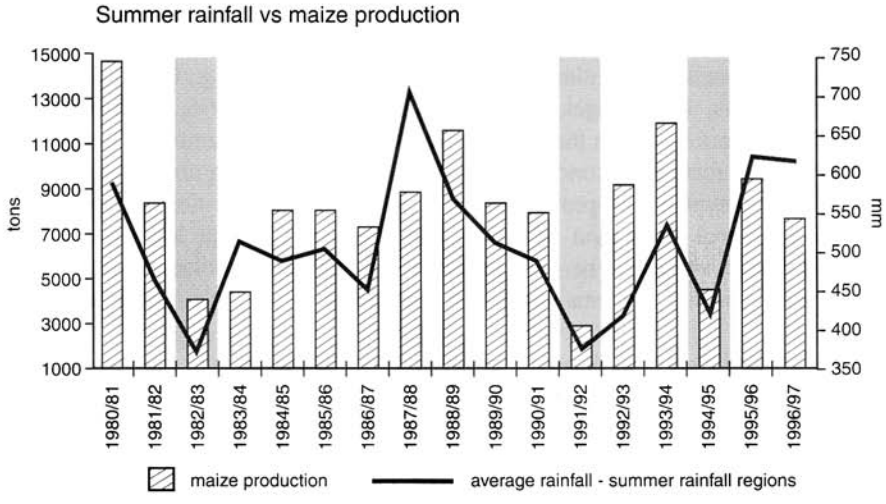


Figure 4 Summer rainfall versus maize production (modified after McClintock, 1997, with kind permission, UBS Warburg, formerly SBC Warburg, Johannesburg).

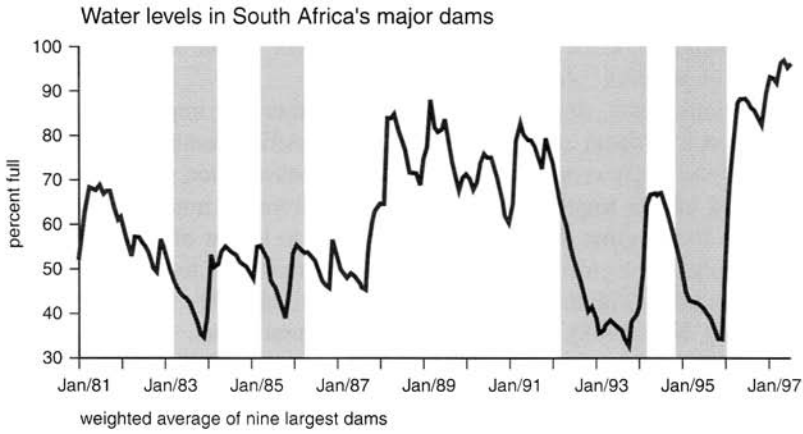


Figure 5 Water levels in South Africa's major dams (modified after McClintock, 1997, with kind permission, UBS Warburg, formerly SBC Warburg, Johannesburg).

Drought is not, therefore, the fundamental problem in sub-Saharan Africa. Drought needs to be viewed together with a host of other hazards and realities: including HIV/AIDS, violence and conflict, growing disparities between rich and poor, failing economies, struggles over land, water, and poverty. Drought indeed often merely uncovers the African development crisis and allows glimpses of harsh daily realities. At a local scale, the impacts of droughts are often hidden "costs" escaping detailed quantification. These include the stripping of household assets used to procure a livelihood (stock and crops are reduced, the price of water and of basic food supplies often increases, retrenchments occur), household income, and social dislocation and disruption of local livelihood (e.g., Bratton, 1987; Adams, 1993; Vogel, 1995; Scoones et al., 1996).

Several assessments of vulnerability to droughts in Africa (e.g., Glantz and Katz, 1985; Chambers, 1989; Vogel, 1995; Jallow, 1995; Davies, 1996; Downing et al., 1996) have therefore shown that droughts act together with a number of underlying factors to exacerbate local conditions. In southern Africa, during non-drought years, the baseline prevalence of problems related to inadequate nutrition is "normally" low (International Federation of Red Cross and Red Crescent Societies, 1999). Drought cannot, therefore, be seen as the "cause" of such problems, but rather, it exacerbates existing problems associated with poverty (e.g., Abrams et al., 1992; AFRA, 1992).

Few detailed studies of drought impacts, coping responses, and mitigation at the rural-poor household level have been undertaken in South Africa (e.g., Freeman, 1984; AFRA, 1992; Vogel, 1995). Most assessments show that it is "access" and "entitlements" to resources that usually determine the magnitude of impact. Rural communities that farm and depend on the land for a livelihood often require access to irrigation, access to markets to sell stock, financial resources to procure farming equipment, etc. Failure to procure these resources, together with a severe dry spell,

can result in inexorable difficulties. During the drought of the early 1990s, for example, large numbers of cattle died in several rural areas [an estimated 500,000 cattle in the former Transkei, a former "independent homeland" in the eastern part of the country (Adams, 1993)]. Such losses of cattle do not only result in less meat and milk but also severely constrain the limited household incomes derived from cattle sales. Underpinning these circumstances is the complex history of the country, which has had a major influence on who farms, owns land, and can obtain access to the resources mentioned above (Lipton et al., 1996).

7 DROUGHT MANAGEMENT AND POLICY INITIATIVES

Droughts, as shown here, are a regular feature in the tapestry of South African history but have been traditionally managed from an agricultural and conservation perspective (Union of South Africa, 1923). This focus has, however, been expanded during the past decade to include a wider group of affected communities and stakeholders. During the droughts of the 1990s, the impact of the drought on rural populations, for example, was actively monitored by various task forces that emerged from a National Consultative Committee on Drought as a result of reports of severe impacts on rural communities in the country, particularly those that had been relocated during the years of apartheid. The activities of the Drought Forum raised the profile and plight of the rural poor during droughts in South Africa and ushered in a change in drought policy (Abrams et al., 1992; Adams, 1993).

Building on the experiences of the 1990s drought, a strong mitigation focus for droughts has been fostered and is contained in the new disaster management policy, which includes a focus on the biophysical resources of the country as well as concentrating on mitigating the host of other factors that exacerbate drought impacts (*White Paper on Disaster Management*, 1999 and Bill, (forthcoming 2002)). The national policy on disasters, including drought, calls for a more proactive response to future droughts in the country. The growing official awareness and policy efforts are envisaged to lead to greater coordination with FEWS (Famine Early Warning Systems) in the wider SADC region and hence improve drought management, both locally and in the region.

Past and present experiences with the vagaries of weather and climate are prompting concerted efforts to improve preparedness for droughts. Collaborative efforts by both forecasters, atmospheric modelers and users of climate outlooks and forecasts have resulted in the formation of various climate forums that have been established throughout Africa, through consensus to improve the quality of the forecasts for the forthcoming seasons. Combining their knowledge of climate conditions in southern Africa, these experts provide users with a consensus, probabilistic assessment of the upcoming rainy season (NOAA-OGP, 1999). The South African Weather Services is actively involved in the forum and through its Research Group for Seasonal Climate Studies, three monthly mean rainfall and temperature forecasts for the country, regularly updated, are produced and issued by the Long-term Operational Group Information Centre (LOGIC). The science of forecasting seasonal rainfall and

temperature for southern Africa and South Africa is well developed (e.g., Joubert et al., 1996; Mason et al., 1996; Joubert and Hewitson, 1997; Mason and Joubert, 1997; Mason, 1997; Mason and Jury, 1997). The overwhelming need, however, still remains for integrated science that incorporates the human and physical dimensions of climate variability and change so that effective mitigation strategies can be initiated and implemented.

Despite these advances, several groups, particularly the rural poor in Southern Africa, remain food insecure. One solution to avoid such situations is to ensure that forecasts are made more accessible. Others, however, argue that much more is required. The eradication of food emergencies and in cases, famines, requires more than technical capacity. Substantial political will, at national and international levels, more than has been evident to date, is needed (Devereux, 2000).

8 CONCLUSIONS

Drought is endemic to the southern African region. Changes in sea surface temperatures together with realignments of pressure systems can, in some cases, trigger a severe drought period such as occurred in the early 1980s and the 1990s. The impacts of such droughts usually result in severe constraints on food production at regional, national, and local levels impacting on GDP, commercial food supply, and water availability. Reductions in water aggravate such problems creating ripple effects that touch on several industries, activities, and communities.

Drought, however, has many facets and it is often the poorly documented cases, such as household vulnerability to drought in poor rural and urban areas, that require much more careful research. Research and understanding of the multidimensional nature of drought, the complex coping strategies in the face of drought, and mitigation strategies are required if effective mitigation and management of droughts is to occur. Research that has been conducted and available vulnerability assessments indicate that drought often unveils many "everyday" realities that are rooted in other factors such as poverty, development, and the complex interlinkages among economic, social, political, and environmental issues. Land degradation, urban and periurban growth, and the impact of HIV/AIDS are some of the factors that heighten vulnerability to the vagaries of weather and climate in Africa. The response to such problems therefore has to be a multifaceted one in South Africa and elsewhere in Africa.

REFERENCES

Abrams, L., Drought policy—water issues, The African Water Page, http://www.african-water.org.drghtwater.htm 1997.

Abrams, L., R. Short, and J. Evans, Root cause and relief restraint report, National Consultative Forum on Drought, Secretarial and Ops Room, Johannesburg, October 8, 1992.

Adams, L., A rural voice, strategies for drought relief, *Indicator S. Afr.*, *10*(4), 41–46, 1993.

Alexander, W. J. R., Floods, droughts and climate change, *S. Afr. J. Sci.*, 91, 403–408.

Allan, R. J., J. A. Lindesay, and D. Parker, *El Nino Southern Oscillation and Climatic Variability*, CSIRO Publishing, Melbourne, Australia, 1996.

Association for Rural Advancement (AFRA), *Drought Relief and Rural Communities*, Special Report No. 9, AFRA, Pietermaritzburg, 1992.

Ballard, C., Drought and economic disasters: South Africa in the 1980s, *J. Interdiscipl. Hist.*, *17*, 359–378, 1986.

Benson, C., and E. J. Clay, Developing countries and the economic impacts of natural disasters, in Kreimer, A., and Arnold, M. (Eds.), *Managing Risk in Emerging Economies*, The World Bank, Washington, D.C., 11–21, 2000.

Bhalotra, Y. P. R., *The Drought of 1981–1985 in Botswana*, Department of Meteorological Services, Ministry of Works and Communications, Gaborone, Botswana, 1985.

Bratton, M., Drought, food and social organization of small farmers in Zimbabwe, in M. Glantz (Ed.), *Drought and Hunger in Africa: Denying Famine a Future*, Cambridge University Press, Cambridge, 1987, pp. 213–244.

Bruwer, J. J., Drought policy in the Republic of South Africa, Part I, *Drought Network News*, *1*(3), 14–16, University of Nebraska, Lincoln, USA, 1989.

Bruwer, J. J., Drought policy in the Republic of South Africa, in A. L. Du Pisani, (Ed.), *Proceedings of the SARCCUS Workshop on Drought*, June, 1989, SARCCUS, Pretoria, 1990, pp. 23–38.

Cane, M. A., G. Eshel, and R. W. Buckland, Forecasting Zimbabwean maize yield using eastern equatorial pacific sea surface temperature, *Nature*, *370*, 204–205, 1994.

Chambers, R., Vulnerability, coping and policy in *IDS Bulletin, 20, 2: Vulnerability: How the Poor Cope*, Institute of Development Studies IDS, University of Sussex, Brighton, England, 1989, pp. 1–7 .

Davies, S., *Adaptable Livelihoods: Coping with Food Insecurity in the Malian Sahel*, Macmillan, London, 1996.

Dent, M. C., R. E. Schulze, H. M. M. Wills, and S. D. Lynch, Spatial and temporal analysis of the recent drought in the summer rainfall region of Southern Africa, *Water SA*, *13*, 37–42, 1987.

Department of Water Affairs and Forestry, *Overview of Water Resources, Availability and Utilization in South Africa*, Department of Water Affairs and Forestry, Pretoria, 1997.

Dilley, M., Climate Change and Disasters, in Kreimer, A., and Arnold, M. (Eds.), *Managing Disaster Risk in Emerging Economies*, The World Bank, Washington D.C., 45–50, 2000.

Downing, T., M. Watts, and H. Bohle, Climate change and food insecurity: Toward a sociology and geography of vulnerability, in T. E. Downing, (Ed.), *Climate Change and World Food Security*, Nato ASI Series, Global Environmental Change, 1996, pp. 183–206.

Du Pisani, A. L., The CERES-MAIZE model as a potential tool for drought assessment in South Africa, *Water SA*, *13*, 159–164, 1987.

Enfield, D. B., El Nino, past and present, *Rev. Geophys. 27*, 159–187, 1989.

Erasmus, J. F., Drought monitoring: Using rainfall DECILES as a drought index, *Pixels and Bytes*, *5*, 42–48, 1987.

Fox, R., and K. Rowntree (Eds.), *The Geography of South Africa in a Changing World*, Oxford University Press, Cape Town, 2000.

Freeman, C., Drought and agricultural decline in Bophuthatswana, in South African Research Services, (Eds.), *South African Review II*, Ravan Press, Johannesburg, 1984, pp. 284–289.

Glantz, M. H. (Ed.), *Once Burned, Twice Shy? Lessons Learned from the 199798 El Nin˜o*, UNEP, NCAR, United Nations University, WMO, ISOR, United Nations University, Japan, 2001.

Glantz, M. H., and R. W. Katz, Drought as a constraint to development in sub-Saharan Africa, *Ambio*, *14*, 334–339, 1985.

Glantz, M. H., R. Katz, and M. Krenz, (Eds.), *The Societal Impacts Associated with the 1982–83 Worldwide Climate Anomalies*, Environmental and Societal Impacts Group, United Nations Publications Office, New York, 1987.

Goudie, A. (Ed.), *The Encyclopedic Dictionary of Physical Geography*, 2nd ed., Blackwell, Oxford, 1994.

Harrison, M. S. J., The Southern Oscillation, zonal equatorial circulation cells and South African rainfall, in *Preprints of the First International Conference on Southern Hemisphere Meteorology*, American Meterological Society, Boston, 1983, pp. 302–305.

Harrison, M. S. J., A generalized classification of South African summer rain-bearing synoptic systems, *J. Climatol.*, *4*, 547–560, 1984a.

Harrison, M. S. J., The annual rainfall cycle over the central interior of South Africa, *S. Afr. Geograph. J.*, *66*, 46–64, 1984b.

Harrison, M. S. J., Comparison of rainfall time series over South Africa generated from real data and through principals component analysis, *J. Climatol. 4*, 561–564, 1984c.

Harrison, M. S. J., A synoptic climatology of South African rainfall variations, Ph.D. thesis, University of the Witwatersand, 1986.

Hulme, M. (Ed.), Climate Change and Southern Africa: An exploration of some potential impacts and implications in the SADC region, Report commissioned by the WWF International and co-ordinated by the Climate Research Unit, UEA, Norwich, United Kingdom, 1996.

International Federation of Red Cross and Red Crescent Societies, (IFRCRCS), *World Disasters Report*, IFRCRCS, Geneva, Switzerland, 1999

Jallow, S. S., Identification of and response to drought by local communities in Fulladu West District, The Gambia, *Singapore J. Trop. Geogr.*, *16*, 22–41, 1995.

Joubert, A. M., and B. Hewitson, Simulating present and future climates of southern Africa using General Circulation Models, *Prog. Phys. Geogr.*, *21*, 51–78, 1997.

Joubert, A. M., S. J. Mason, and J. S. Galpin, Droughts over southern Africa in a doubled-CO_2 climate, *Int. J. Climatol. 16*, 1149–1156, 1996.

Jury, M. R., A review of research on ocean-atmosphere interactions and South Afrian climate variability, *S. Afr. J. Sci.*, *91*, 289–294, 1995.

Jury, M. R., and K. M. Levey, The climatology and characteristics of drought in the Eastern Cape of South Africa, *Int. J. Climatol.*, *13*, 629–641, 1993.

Jury, M. R., B. M. R. Pathack, C. J de W. Rautenbach, and J. van Heerden, Drought over South Africa and Indian Ocean SST: Statistical and GCM results, *Global Atmos. Ocean Syst. 4*, 47–63, 1996.

Laing, M., Drought update 1991–1992, South Africa, *Drought Network News*, *4*(2), 15–17, University of Nebraska, Lincoln, USA, 1992.

Landman, W. A., and S. J. Mason, Change in the association between Indian Ocean sea-surface temperatures and summer rainfall over South Africa and Namibia, *Int. J. Climatol.*, *19*, 1477–1492, 1999.

Lindesay, J. A., Southern African rainfall, the Southern Oscillation and a Southern Hemisphere semi-annual cycle, *J. Climatol.*, *8*, 17–30, 1988.

Lindesay, J. A., Present climates of southern Africa, in J. E., Hobbs, J. A. Lindesay, and H. A Bridgman (Eds.), *Climates of the Southern Continents Present, Past and Future*, Wiley, Chichester, 1998.

Lindesay, J. A., and C. H. Vogel, Historical evidence for Southern Oscillation–southern African rainfall relationships, *Int. J. Climatol.*, *10*, 679–689, 1990.

Lipton, M., F. Ellis, and M. Lipton, *Land, Labour and Livelihoods in rural South Africa*, Vol. 2: *Kwa Zulu Natal and Northern Province*, Indicator Press, University of Natal, South Africa, 1996.

Mason, S. J., Temporal variability of sea surface temperatures around southern Africa: A possible forcing mechanism for the eighteen-year rainfall oscillation? *S. Afr. J. Sci.*, *86*, 243–252, 1990.

Mason, S. J., Review of recent developments in seasonal forecasting of rainfall, *Water SA*, *23*, 57–62, 1997.

Mason, S. J., L. Goddard, N. E. Graham, Yulaeva, L. Sun, and P. A. Arkin, The IRI seasonal climate prediction system and the 1997/98 El Nino event, *Bull. Am. Meteorol. Soc.*, *80*, 1853–1973, 1999.

Mason, S. J., and M. R. Jury, Climatic variability and change over southern Africa: a reflection on underlying processes, *Prog. Phys. Geogr.*, *21*(1), 23–50, 1997.

Mason, S. J., and A. M. Joubert, Simulated changes in extreme rainfall over southern Africa, *Int. J. Climatol.*, *17*, 291–301, 1997.

Mason, S. J, A. M. Joubert, Cosijn, C. and S. J. Crimp, Review of the current state of seasonal forecasting techniques with applicability to Southern Africa, *Water SA*, *22*(3), 203–209, 1996.

Mason, S. J., J. A. Lindesay, and P. D. Tyson, Simulating drought over southern Africa using sea surface temperature variations, *Water SA*, *20*, 15–21, 1994.

Mason, S. J., and P. D. Tyson, The occurrence and predictability of droughts over southern Africa, in D. Wilhite, (Ed.), *Drought*, Vol. 1: *A Global Assessment, Routledge Hazards and Disasters Series*, Routledge, 2000, pp. 112–134.

Mather, C., and A. Adelzadeh, Macroeconomic strategies, agriculture and rural poverty in post-apartheid South Africa, paper presented for the Land and Agricultural Policy Centre, Johannesburg, 1997.

McClintock, M., *El Nino, Cloud on the Horizon, Economics, SBC Warburg*, Swiss Bank Corporation, Johannesburg, South Africa, 1997.

Miron, O., and J. A. Lindesay, A note on changes in airflow patterns between wet and dry spells over South Africa, 1963 to 1979, *S. Afr. Geogr. J.*, *65*, 141–147, 1983.

Muller, M. J., and P. D. Tyson, Winter rainfall over the interior of South Africa during extreme dry years, *S. Afr. Geogr. J.*, *70*, 20–30, 1988.

National Oceanic and Atmospheric Administration, Office of Global Programs (NOAA-OGP), U.S. Department of Commerce, An experiment in the application of climate forecasts: NOAA-OGP activities related to the 1997-98 El Nino event, Washington, D.C., NDAA/OGP, 1999.

Ngara, T., D. L. McNaughton, and S. Lineham, Seasonal rainfall fluctuations in Zimbabwe, *Zimbabwe Agri. J.*, *80*, 149–150, 1983.

Nicholson, S. E., Climatic variations in the Sahel and other African regions during the past five centuries, *J. Arid Environ.*, *1*, 3–24, 1978.

Nicholson, S. E., African drought: Characteristics, causal theories and global teleconnections, in A. Berger, R. E. Dickinson, and J. W. Kidson (Eds.), *Understanding Climate Change, Geophysical Monograph*, *52*, American Geophysical Union, Washington, DC, 1989, pp. 79–100.

Nicholson, S. E., and D. Entekhabi, The quasi-periodic behaviour of rainfall variability in Africa and its relationship to the Southern Oscillation, *Arch. Meteorl. Geophys. Bioklimatol. Ser. A.*, *34*, 311–348, 1986.

Ogallo, L. J., Impacts of the 1982–83 ENSO event on eastern and southern Africa, in M. H. Glantz, R. Katz, and M. E. Krenz (Eds.), *The Societal Impacts Associated with the 198283 Worldwide Climate Anomalies*, United Nations Publications Office, New York, pp. 55–61, 1987.

Ogallo, L. J., J. E. Janowaik, and M. S. Halpert, Teleconnections between seasonal rainfall over East Africa and global sea surface temperature anomalies, *J. Meteorol. Soc. J.*, *66*, 807–821, 1988.

Partridge, T. C., Cainozoic environmental change in southern Africa, with special emphasis on the last 200 000 years, *Progr. Phys. Geogr.*, *21*(1), 3–22, 1997.

Philander, S. G. H., Unusual conditions in the tropical Atlantic Ocean in 1984, *Nature*, *322*, 236–238, 1986.

Schulze, R. E., Hydrological simulation as a tool for agricultural drought assessment, *Water SA*, *10*, 55–62, 1984.

Schulze, R. E., Modelling hydrological responses to land use and climate change: A southern African perspective, *Ambio*, *29*, 12–22, 2000.

Schulze, R., J. Meigh, and M. Horan, Present and future vulnerability of eastern and southern Africa's hydrology and water resources, *S. Afr. J. Sci.*, *97*, 150–160.

Scoones, I., C. Chibudud, S. Chikara, P. Jeranyama, D. Machaka, W. Machanja, B. Mavedzenge, B. Mombeshpra, M. Mudhara, C. Mudsiwo, F. Murimbarimba, and B. Zirera, *Hazards and Opportunities: Farming Livelihoods in Dryland Africa: Lessons from Zimbabwe*, Zed Press, London, 1996.

Taljaard, J. J., Change of rainfall distribution and circulation patterns over southern Africa in summer, *J. Climatol.*, *6*, 579–592, 1986.

Taljaard, J. J., *Climate and Circulation Anomalies in the South African Region during the Dry Summer of 19824983* , South African Weather Bureau Technical Paper No. 21, Weather Bureau, Pretoria, 1989.

Thomson, A., P. Jenden, and E. Clay, Information, risk and preparedness: Responses to the 1997 El Nino event, Research report, DFID, ESCOR No AG1215, SOS SAHEL, London, 1998.

Torrance, J. D., Malawi, Rhodesia and Zambia, in J. F. Griffiths (Ed.), *Climates of Africa, World Survey of Climatology*, Vol. 10, Elsevier, Amsterdam, 1972, pp. 409–460.

Tyson, P. D., Atmospheric circulation variations and the occurrence of extended wet and dry spells over southern Africa, *J. Climatol.*, *1*, 115–130, 1981.

Tyson, P. D., The atmospheric modulation of extended wet and dry spells over South Africa, 1958–1978, *J. Climatol. 4*, 621–635, 1984.

Tyson, P., *Climate Change and Variability in Southern Africa*, Oxford University Press, Cape Town, 1986.

Tyson, P. D., and T. G. J. Dyer, The predicted above-normal rainfall of the seventies and the likelihood of droughts in the eighties in South Africa, *S. Afr. J. Sci.*, *74*, 372–377, 1978.

Tyson, P. D., and T. G. J. Dyer, 1980: The likelihood of droughts in the eighties in South Africa, *S. Afr. J. Sci.*, *76*, 340–341, 1980.

Tyson, P. D., T. G. J. Dyer, and M. N. Mametse, Secular changes in South African rainfall: 1880 to 1972, *Q. J. Roy. Meteorol. Soc.*, *101*, 817–833, 1975.

Tyson, P. D., and J. A. Lindesay, The climate of the last 2000 years in southern Africa, *Holocene*, *2*, 271–278, 1992.

Tyson, P. D., and R. A. Preston-Whyte, *The Weather and Climate of Southern Africa*, Oxford University Press, Cape Town, 2000.

Union of South Africa, *Final Report of the Drought Investigation Commission*, UG 49–23, Government Printers, Cape Town, 1923.

Van Zyl, J., The last straw: Drought and the economy, *Indicator SA*, *10*(4), 47–51, 1993.

Van Zyl, J., and H. J. G. Nel, The role of the maize industry in the South African economy, *Agrekon*, *27*, 10–16, 1988.

Van Zyl, J., A. Van der Vyver, and J. A. Groenewald, The influence of drought and general economic effects on agriculture: A macro-analysis, *Agrekon*, *26*, 8–12, 1987.

Vogel, C., A documentary-derived chronology for southern Africa, 1820–1900, *Climate Change*, *14*, 291–306, 1989.

Vogel, C. H., People and drought in South Africa: reaction and mitigation, in T. Binns (Ed.), *People and Environment in Africa*, Wiley, London, 1995, pp. 249–256.

Vogel, C. H., M. Laing, and K. Monnik, Drought in South Africa, with special reference to the 1980–1994 period, in D. Wilhite (Ed.), *Drought*, Vol. 1: *A Global Assessment, Routledge Hazards and Disasters Series*, Routledge, London, 2000, pp. 348–366.

Walker, N., Links between South African summer rainfall and temperature variability of the Agulhas and Benguela currents systems, *J. Geophys. Res.*, *95*, 3297–3319, 1990.

Walker, N., J. Taunton-Clark, and J. Pugh, Sea temperatures off the South African west coast as indicators of Benguela warm events, *S. Afr. J. Sci.*, *80*, 72–77, 1984.

Walker, N. D., and J. A. Lindesay, Preliminary observations of oceanic influences on the February–March 1988 floods in central South Africa. *S. Afr. J. Sci.*, *85*, 164–169, 1989.

Wilhite, D. A., The role of government in planning drought: Where do we go from here? in D. A. Wilhite, W. E. Easterling, and D. A. Woods (Eds.), *Planning for Drought: Toward a Reduction of Societal Vulnerability*, Westview Press, Boulder, CO, 1987, pp. 425–444.

Wilhite, D. A., and M. H. Glantz, Understanding the drought phenomenon: The role of definitions, *Water Int.*, *10*, 111–120, 1985.

White Paper on Disaster Management, Ministry for Provincial Affairs and Constitutional Development, Pretoria, 1999.

World Food Programme, WFP, Hunger in Southern Africa: The Unfolding Crisis, 30-05-2002, http://www.wfp.org/index

Zucchini, W. Z., and P. T. Adamson, The occurrence and severity of droughts in South Africa, report to the Water Research Commission, WRC 91/1/84, prepared by the Department of Civil Engineering, University of Stellenbosch and the Department of Water Affairs, Pretoria, 1984.

CHAPTER 46

TRANSBOUNDARY FISHERIES: PACIFIC SALMON

KATHLEEN A. MILLER AND MARY W. DOWNTON

1 INTRODUCTION

Climate variations often affect the abundance, location, or migratory patterns of fish populations. Even when a fishery is entirely contained within a single jurisdiction, these climate impacts complicate the difficult task of maintaining economically efficient and biologically sound harvest management while balancing the interests of competing harvesters. When fish stocks are harvested by more than one nation, or when they cross internal jurisdictional boundaries, the management task is further complicated by the efforts of each nation or jurisdiction to promote the interests of its own harvesters.

The faltering attempts of the United States and Canada to cooperate on Pacific salmon management illustrate the fragile nature of such cooperation and the destabilizing role that climate variations can play. These two nations have a long and rocky history of alternating between cooperative salmon conservation efforts and predatory grabs at one another's returning adult salmon. The most recent breakdown in cooperation began in 1993. For 6 years, the United States and Canada were unable to agree on a full set of salmon "fishing regimes" under the terms of the Pacific Salmon Treaty. The conflict was sparked by strongly divergent trends in the abundance of northern and southern salmon stocks and a consequent change in the balance of each nation's interceptions of salmon spawned in the other nation's rivers. Alaska's salmon harvests (i.e., northern) have experienced a remarkable sustained increase over the past two decades, while harvests of some salmon stocks in British Columbia and chinook and coho harvests in Washington, Oregon and California (i.e., southern) have fared poorly. These trends appear to be influ-

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

enced by the impacts of climatic variations on stock abundance, but climate is not the only source of harvest variability. Because it is not easy to disentangle natural and anthropogenic sources of variability, the negotiation process has been complicated by differences of opinion over the biological "facts."

A new agreement, signed on June 30, 1999, may end the conflict, but it is too early to judge its likelihood of success. The Canadians remain bitterly divided over the merits of the agreement, which has been labeled a "sellout" by Canadian fishing interests, and the arrangement is still contingent on U.S. Congressional approval of $140 million for two jointly managed endowment funds to be used for scientific cooperation, stock enhancement, and habitat restoration (Culbert and Beatty, 1999).

This case exemplifies many of the problems that arise in the management of transboundary fishery resources. Each nation has the power to significantly damage the other's interests and, in the short term, each could gain by interfering with the other's harvests. In the longer term, such opportunistic competition tends to squander the potential value of the fishery, as each nation commits excessive fishing capital and labor in its attempts to capture a larger share of the available fish. The ultimate result of competitive harvesting may be a "tragedy of the commons" (Hardin, 1968) in which overfished stocks are depleted and the fishery declines, sometimes abruptly.

Nations that exploit shared fishery resources often recognize the mutually destructive effects of unconstrained competitive harvesting, and they may attempt to improve the situation by negotiating harvest management agreements. The stability of such agreements hinges on the extent to which it is easy or difficult to monitor and enforce compliance and on the extent to which each party continues to expect to gain by cooperating. Many such agreements have proven to be unstable. The tendency for cooperation to degenerate into mutually destructive fish wars is a significant puzzle. The problem may have its roots in high costs of monitoring and enforcement as well as in the fact that the parties' incentives to cooperate change over time (Miller, 1996; McKelvey, 1997). Climatic variations contribute to these sources of instability by causing fluctuations in fish abundance and availability that are difficult to observe and predict. One can see the playing-out of this process in the collapse of cooperation in the Pacific Salmon Treaty case.

2 SALMON ABUNDANCE: CLIMATE AND OTHER INFLUENCES

Pacific salmon are anadromous fish that spawn in streams from California's Central Valley northward. In spring, juvenile salmon emerge from the freshwater environment and disperse into the coastal ocean. Some salmon stocks remain in coastal areas throughout their lives, but many others spend a year or more in a long-distance migration across the feeding grounds of the subarctic Pacific before returning to their natal streams to spawn and to die (Pearcy, 1992). There are five species of Pacific salmon, with a multitude of distinct breeding populations. All five species (chinook, coho, sockeye, pink, and chum) are present from Washington state northward, while in Oregon and California only chinook and coho spawn in significant numbers.

In the mid-1970s, ocean conditions in the North Pacific changed dramatically. Shortly thereafter, Alaskan salmon harvests entered a period of dramatic increase, rising nearly 10-fold from a low of 22 million salmon (of all species) in 1974 to three successive record highs in 1993, 1994 and 1995 (Fig. 1). At the 1995 peak, Alaska harvested a total of 217 million salmon. Harvests of most salmon species in northern British Columbia also fared well during this period, although British Columbia's commercial chinook harvests have declined steadily, and by the late 1990s it became apparent that many of British Columbia's southern and interior coho stocks are severely depleted. Southward, salmon harvests have been on a roller-coaster. Commercial chinook and coho catches in California, Oregon, and Washington dropped abruptly in the late 1970s, hitting El Niño–related lows in 1983 and 1984. A dramatic but brief recovery in 1986 and 1987 then gave way to a precipitous decline to record low harvests in recent years (Fig. 2). Production has declined to the point that some stocks are on the verge of extinction. Some observers attribute these changes in salmon productivity to human disruptions of the southern fisheries and good management of the northern ones (Royce, 1988). However, mounting evidence suggests that shifts in marine climate may have played a major role (Beamish and Bouillon 1993; Hare and Francis 1995; Mantua et al., 1997).

Winter climate over the North Pacific is dominated by a low-pressure system centered near the Aleutian Islands. From 1977 through 1988, the Aleutian Low was frequently deeper than normal, leading to severe storms, increased mixing, and cooler temperatures in the central North Pacific (Trenberth and Hurrell, 1994) (Fig. 3). Along the west coast of North America, sea surface temperatures (SST) were unusually warm during the 1977 to 1995 period (Fig. 4).

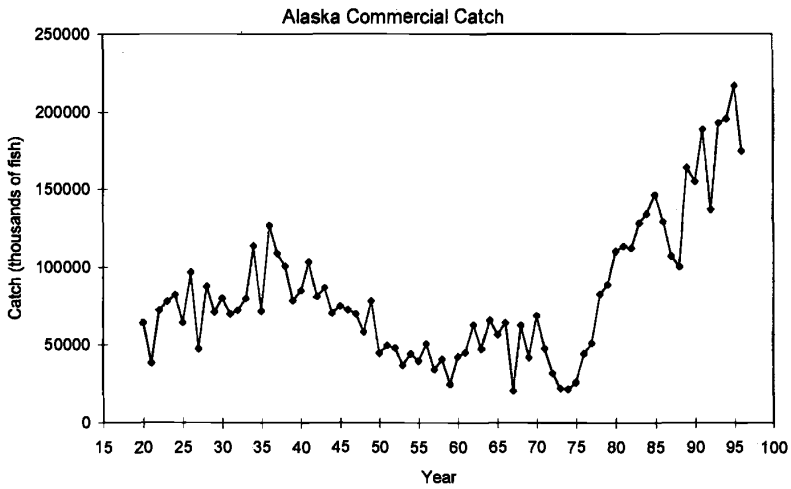


Figure 1 Alaskan commercial salmon harvest—all species.

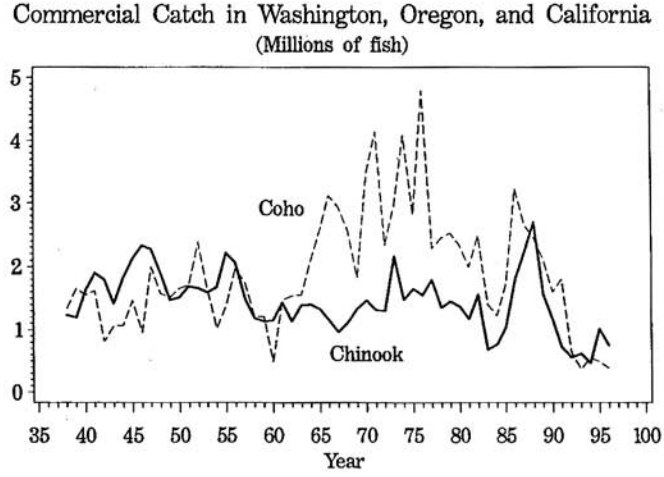


Figure 2 Commercial coho and chinook Salmon harvest in Washington, Oregon, and California, millions of fish.

Possible causes of the change are the subject of much research. In the last century, the North Pacific climate varied on an interdecadal scale, with shifts or trends in mean levels of sea-level pressure and SST that lasted for several decades (Zhang et al., 1996; Latif and Barnett, 1996). The pattern is characterized by alternate warm

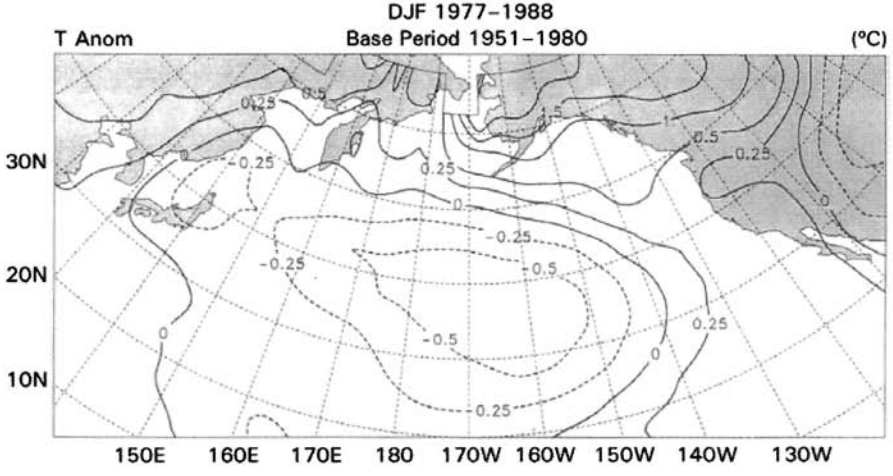


Figure 3 Twelve-year (1977–1988) average winter surface temperature anomalies (°C) shown as departures from the 1951–1980 mean. Gridded temperature data consists of air temperatures over land and sea surface temperatures over oceans (IPCC 1992; Trenberth et al., 1992). Figure courtesy of James Hurrell.

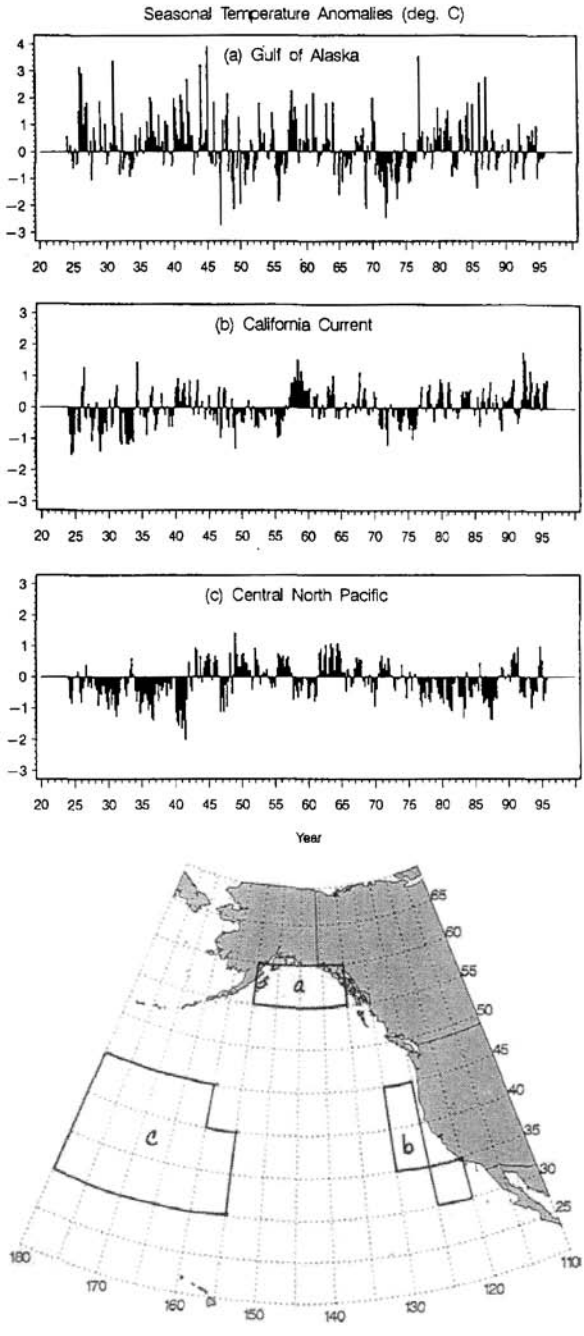


Figure 4 Seasonal surface temperature anomalies in three regions of the northeast Pacific: (*a*) Gulf of Alaska, (*b*) California current, and (*c*) central North Pacific. (Temperature data as described in Fig. 3.)

and cool periods in a large area of the western and central north Pacific, with shifts toward warmer temperatures in the mid-1940s and cooler temperatures in the mid-1970s (the eastern edge of this region is area C in Fig. 4). The cool periods in the central North Pacific are associated with an intensification of the Aleutian Low and a warming of coastal temperatures along the west coast of North America.

The North Pacific also is influenced by the El Niño–Southern Oscillation (ENSO) phenomenon (Kiladis and Diaz, 1989). ENSO-related warming of the equatorial Pacific occurs intermittently, at intervals of about 3 to 6 years, and frequently leads to intensification of the Aleutian Low. The effects of an ENSO warm event often propagate northward, warming the west coast of North America and cooling SSTs in the central north Pacific. An unusual sequence of closely spaced ENSO warm events have occurred from 1977 to 1995, possibly evidence of a change in large-scale climate (Trenberth and Hurrell, 1994; Trenberth and Hoar, 1996). These ENSO warm events (El Niño events) have tended to reinforce the decadal-scale shift to warmer coastal SSTs and cooler SSTs in the central north Pacific.

Intensification of the Aleutian low and warming of the coastal ocean appear to have positive effects on salmon abundance in the Gulf of Alaska, but negative effects on stocks that spend a portion of their lives in the California current (Pearcy, 1992; Hare et al., 1999). In the subarctic zone, the mixed layer has become shallower. This may have enhanced the survival of Alaskan and northern British Columbian salmon smolts by increasing the productivity of zooplankton, which is a frequent food source for juvenile salmon (Polovina et al., 1995; Brodeur and Ware, 1992). A general pattern of winter warming and increased winter precipitation in Alaska over the past two decades (Mantua et al., 1997) also may have contributed to favorable stream conditions for egg-to-smolt survival. From southern British Columbia southward, El Niño events have been associated with poor feeding conditions for maturing salmon and changes in species composition, including increased abundance of some species that prey on juvenile salmon (Pearcy, 1992). In addition, recent droughts in California and the Pacific Northwest have resulted in poor conditions for spawning and migration in the salmon's freshwater phase. Changes in ocean temperatures and circulation, and associated changes in stream conditions, thus appear to have contributed to the opposite trends in northern and southern salmon production.

3 HISTORY OF HARVEST MANAGEMENT

North America's commercial Pacific salmon fisheries were established and grew rapidly in the late nineteenth century. In many areas, returning adult salmon soon were running such a gauntlet of competing fishing gears and other hindrances, such as dams, that it was a lucky salmon that survived to spawn (Higgs, 1982). The resulting decline in salmon populations led to the creation of public agencies to establish fishing gear restrictions and fishing seasons. However, these authorities could never fully control harvests of the salmon stocks within their purview because many salmon could be caught as they passed through the waters of neighboring

jurisdictions on their return migration. Such "interceptions" became increasingly important over time as fishing effort expanded in offshore areas.

The first significant international agreement on salmon harvests was a convention between the United States and Canada, signed in 1930 and ratified in 1937. That agreement divided the harvest of Fraser River sockeye salmon as well as management and restoration costs equally between the two nations (Munro and Stokes, 1989). The agreement was later extended to Fraser River pink salmon. Under that convention, the International Pacific Salmon Fishery Commission (IPSFC) regulated harvests of the Fraser River stocks. Although the Fraser River lies entirely in Canada, a large portion of the salmon spawning in that drainage typically approach the river through the Strait of Juan de Fuca where, historically, they had been harvested by Washington State fishing vessels. When a rock slide blocked access to part of their spawning habitat, and sent the Fraser's salmon resources into decline, the United States and Canada had a clear joint interest in removing the blockage and restoring the runs.

That agreement covered only a portion of the salmon runs jointly exploited by the United States and Canada. When negotiations for the Pacific Salmon Treaty began in 1971, Alaskan interceptions of salmon spawned in the rivers of Washington and Oregon were creating tensions among the states, while increasing Canadian troll harvests of those stocks precluded an effective internal solution. In addition, mutual interceptions of salmon of Canadian and Alaskan origin were seen as a barrier to effective management in the northern area (Yanagida, 1987). At the same time, the Canadians had become increasingly unhappy about their agreement to share one half of the Fraser River salmon with the United States because, by foregoing construction of hydropower dams on the Fraser, Canada was effectively bearing more than half of the cost of maintaining those runs. After 14 years of negotiations, the treaty went into effect in 1985.

The treaty created the Pacific Salmon Commission and gave it the authority to promulgate "fishing regimes" to govern harvests in six distinct fisheries. The commission is to periodically renegotiate these regimes as they expire. The body of the treaty lays out a set of general principles to guide the commission in this task. Of central importance are the equity and conservation objectives, which the treaty expresses as follows:

> "...each Party shall conduct its fisheries and its salmon enhancement programs so as to:
> a) prevent overfishing and provide for optimum production; and
> b) provide for each Party to receive benefits equivalent to the production of salmon originating in its waters. (Pacific Salmon Treaty, Article III)"

The treaty then advises the parties to consider the following factors: the desirability of reducing interceptions, the desirability of avoiding disruption of existing fisheries, and annual variations in abundances of the stocks. These considerations are somewhat mutually inconsistent because many of the existing fisheries relied heavily on interceptions.

Until the June 1999 amendments to the treaty, the fishing regimes consisted primarily of harvest ceilings for specific locations and species. For example, in 1985 and 1986, the annual all-gear harvest of chinook in northern and central British Columbia and southeast Alaska was to be limited to 526,000 fish divided equally between the parties (*Treaty*, Annex IV, Chapter 3). Although the intent of the treaty was to control interceptions of fish produced in other jurisdictions, the difficulty of identifying the true origins of fish taken in an ongoing mixed-stock fishery led to the harvest ceiling approach as a proxy method for balancing interceptions.

In addition, the regimes were effective for only a few years. Negotiations for new regimes were to follow a consensus rule, but that allowed any of the parties to veto proposed fishing regimes seen as contrary to its constituent's interests (Yanagida, 1987; Miller, 1996; Schmidt, 1996; Munro et al., 1998). The relevant parties in this context are Canada and the 3 voting U.S. commissioners—representing Alaska, Washington/Oregon, and 24 treaty tribes located in Washington, Oregon, and Idaho. While the Canadian federal government has primary authority on the Canadian side, the British Columbia Provincial Government has often differed vociferously with federal policies, and those internal differences frequently colored the course of the negotiations. When the parties failed to agree on fishing regimes, regulatory authority reverted to the appropriate state or federal jurisdiction. In the United States, the states have authority within 3 nautical miles of the coast and federal jurisdiction extends from 3 to 200 miles offshore.

4 RECENT CONFLICT

The recent breakdown in efforts to renegotiate the expired fishing regimes revolved around two issues. The first was a long-standing dispute over the meaning and enforcement of the treaty's equity provisions. The second was disagreement regarding actions required to meet the treaty's stated goal of rebuilding chinook stocks from the Columbia River northward to southeastern Alaska. When the treaty went into effect, all parties recognized that interceptions could not be reduced to zero and that the interception balance would vary from year to year. They also recognized that the balance would tend to favor either the United States or Canada in each of the six covered fisheries. Canada hoped, however, that the treaty would lead to a rough balance in total interceptions. In particular, they expected that their own interceptions of U.S. coho and chinook would roughly offset the value of U.S. interceptions of Fraser River salmon (Munro and Stokes, 1989; Munro et al., 1998).

Nature and the actions of each party to the agreement have thwarted these expectations. The Canadian commissioners charged that the harvest ceilings failed to ensure an equitable division of the catch. They claimed that Alaska consistently intercepted an excessive number of Canadian salmon. Canada was unable to offset increasing Alaskan interceptions because declining southern coho and chinook abundance prevented Canadian harvesters from reaching the agreed-upon ceilings for harvests of those stocks along the west coast of Vancouver Island. At the same time, fishing interests along the U.S. West Coast claimed that Canada's efforts to

reach the ceiling resulted in overharvesting of those fragile stocks. Alaskan officials countered the Canadian charges by arguing that increased interceptions were unavoidable given the increased abundance of their own intermingled stocks.

When the fishing regimes expired, the Canadians used the opportunity to reassert their demands for a quantitative approach to the equity issue. In previous rounds of regime setting, the Canadians had acquiesced to the quota approach. However in their view, the treaty principle that each party should receive "benefits equivalent to the production of salmon originating in its waters" (*Treaty*, Article III, para. 1) should be interpreted literally as a dollar-for-dollar balancing of the value of a nation's harvest with the value of the salmon spawned in its waters. According to Canadian calculations, the United States owed Canada a considerable debt. The U.S. delegation never favored a quantitative approach, arguing instead that: "[A]n effort to create an accounting scheme would invite costly, and perhaps divisive and inconclusive debate over biological and economic variables" (Yanagida, 1987, p. 591). U.S. officials have been quick to point out that slightly different biological assumptions and valuation rules can give vastly different results regarding amounts owed and even the direction of the equity imbalance (Personal communication, Tom Cooney, Washington State Department of Fish and Wildlife, June 22, 1995; Munro and Stokes, 1989). The U.S. side preferred to treat each of the covered fisheries separately, giving due recognition to the treaty provision disfavoring economic disruption of historic fisheries. Each party's refusal to accept the other's approach to the equity issue resulted in a protracted stalemate.

In addition to Canadian/U.S. differences over implementation of the treaty's equity provisions, a rift developed between Alaska and the other U.S. parties over chinook harvests. When the treaty was ratified, the parties agreed to a program of limiting harvests in order to rebuild naturally spawning chinook stocks by the year 1998. While Alaska's chinook harvests have not increased, declining runs in British Columbia and the southern U.S. jurisdictions pushed the rebuilding goal further out of reach. Tensions on the U.S. side reached a boiling point in 1995 when the Northwest treaty tribes and the states of Washington and Oregon sued Alaska and won an injunction that closed the southeastern Alaska chinook fishery for the remainder of the season (*Confederated Tribes and Bands v. Baldridge* [W.D. Wash. September 7, 1995].

As the treaty dispute escalated, the Canadians employed a variety of desperate tactics in an effort to force the United States back to the bargaining table. For example, in 1994, British Columbia tried to pressure the southern U.S. parties by pursuing an "aggressive fishing strategy." That strategy failed to win any concessions and resulted in dangerous overharvesting of part of the Fraser River sockeye run by the Canadian fleet (Fraser River Sockeye Public Review Board, 1995). By the summer of 1997, British Columbia's salmon harvesters had become so frustrated that approximately 150 fishing vessels participated in a blockade that held the Alaska Ferry hostage in the Canadian port of Prince Rupert for 3 days (Hogben et al., 1997; D'oro, 1997).

Canadian frustration was fueled by Alaska's unwillingness to take actions to reduce the interceptions imbalance. Such concessions made little sense from Alas-

ka's standpoint because they would impose costs on Alaska without commensurate benefits. In fact, Alaska never had much to gain by participating in the Pacific Salmon Treaty, and apart from possible suits over interference with Native American treaty fishing rights, the other U.S. parties have very few bargaining chips to use in their negotiations with Alaska. Alaska's favorable position arises from the fact that many salmon stocks swim northward as juveniles to feed and mature in the Gulf of Alaska. As adults, they swim southward to return to their natal streams. This migratory pattern gives Alaska a natural advantage in exploiting chinook salmon from the southern U.S. jurisdictions and certain Canadian stocks while harvesters in the southern U.S. jurisdictions do not intercept Alaskan origin salmon.

In addition, as Alaska's own salmon became more numerous, Alaska's fishery managers found it increasingly difficult to limit interceptions. In southeastern Alaska, salmon harvesting occurs primarily in areas where Alaska's fish are intermingled with stocks originating elsewhere. Fish caught in those offshore areas are in top condition and at peak value. The Alaskans argue that interceptions cannot be kept constant when their own stocks increase, unless they allow a larger number of their own fish to escape harvesting in those prime areas. Those fish could contribute to spawning escapements or they might later be harvested in a river or estuary where their commercial value is often much lower. However, with spawning escapements already strong, and markets glutted with lower valued "canning quality" salmon, neither of those options is attractive. The Alaskan stance in the negotiations has consistently been that they should be allowed to reap the rewards of their own good salmon management and that they are not to blame for the declining southern stocks.

Canadian efforts to promote a fish-for-fish approach to the equity issue, Alaskan intransigence, and the helplessness of the other U.S. parties to halt the declines of their salmon resources collided in a dangerous muddle. This situation threatened the health of some highly valued salmon stocks and cast a rancorous cloud on relations between the two nations.

5 CURRENT AGREEMENT AND PROSPECTS FOR THE FUTURE

The 1999 agreement represents a dramatic break from the previous approach. Rather than relying on short-lived, ceiling-based regimes whose frequent renegotiation provided ample opportunity for disagreement and brinkmanship, the new agreement establishes a long-term commitment to define harvest shares as a function of the abundance of each salmon species in the areas covered by the treaty. For example, for the next 12 years, the U.S. share of Fraser River sockeye will be fixed at 16.5% of the annual harvest. This represents a decrease from the post-1985 average U.S. share of 20.5%, but an increase relative to the share actually attained by the U.S. fleet during the 1992 to 1997 salmon war period (DFO, 1999; O'Neil, 1999a). This percentage approach allows the number of Fraser River sockeye harvested by the U.S. fleet to increase in years of high sockeye abundance while requiring reduced harvests when abundance is depressed. In contrast, in the 1985 treaty, U.S. harvests

of Fraser sockeye were to be held to a cap of 7 million fish over each of two successive 4-year periods (*Pacific Salmon Treaty*, Annex 4).

The new arrangements for chinook, which will be in effect for 10 years, take account of the fact that the various fisheries along the coast differ considerably in the extent to which they rely on healthy or depressed chinook stocks (U.S. Department of State, 1999). Accordingly, the agreement designates two types of fisheries: (1) aggregate abundance-based management (AABM) fisheries will be managed based on indices of the aggregate abundance of chinook present in the fishery and (2) individual stock-based management (ISBM) fisheries, which are primarily located in inside fishing areas, will be managed based on the status of individual stocks or groups of stocks (e.g., on the basis of the evolving status of currently endangered or threatened stocks). Abundance-based allocation rules for coho have not yet been developed, but the agreement instructs the parties to jointly develop such a management approach, and specifies various deadlines for the accomplishment of particular tasks.

It is not yet clear if the new approach will provide a path to sustainable cooperation. While the focus on conservation will tend to protect some of the weak stocks that were jeopardized by the recent turmoil, the new agreement does little to resolve long-standing differences over the division of benefits. In particular, many Canadians remain convinced that Canada will come out "short" under this agreement, and they have labeled it "profoundly disappointing" (Culbert and Beatty, 1999). They are also unhappy about the long-term implications of the agreement because they feel that it risks committing them to an unsatisfactory arrangement for many years. In fact, the Canadian delegation had unsuccessfully argued for a shorter term. Although the parties have formally stipulated that compliance with the terms of the new agreement shall be deemed to fulfill the requirements of Article III of the treaty, the stipulation applies only for the duration of the current agreement. If Canadians continue to feel that their interests have been compromised, there may be renewed turmoil when this agreement expires. Nevertheless, the agreement represents a step in the right direction and may perhaps lay the groundwork for more robust future cooperation.

One positive aspect is the agreement's provision for two endowment funds, the proceeds of which are to be used to support scientific research, habitat restoration, and enhancement of wild stock production in their respective (northern and southern) areas. Initially, the funding is to be provided entirely by the United States, and the entire agreement is contingent on U.S. Congressional approval of $75 million for the Northern Fund and $65 million for the Southern Fund. In the future, either party may make additional contributions, and even third parties may contribute, with the consent of the parties.

Some analysts have suggested that expanding the scope of the salmon negotiations to allow "side payments" might provide a path to sustained cooperation. These payments might be either in monetary form or in the form of concessions on other nonsalmon issues (Schmidt 1996; Munro et al., 1998). The endowment funds might be a vehicle for providing such side payments, although their initial yield will be far smaller than the debt claimed by Canada from the United States for the accumulated

harvest imbalance. A portion of the money available from the Northern Fund will support Alaskan research and enhancement (O'Neil, 1999b), which, together with other payments, might elicit greater cooperation from Alaska. However, it remains to be seen if this tool will be put to good use, or if its promise will be squandered in quarreling over distribution of the proceeds.

In summary, the new agreement is a hopeful sign, but it does not ensure sustained cooperation. The division of benefits is still largely tied to the division of the catch. Depending on the vagaries of nature, the parties may or may not find that division to be satisfactory. If they do not, and if they fail to develop other methods to ensure a fair division of benefits, then another breakdown in cooperation will likely occur when this agreement expires. The Pacific salmon case demonstrates that the societal impacts of weather and climate are often a complex product of physical or biological impacts, institutional factors, and economic motivations operating in a context of incomplete information. We must attempt to understand all aspects of this complex interaction if we are to improve society's ability to cope with the effects of climatic variations.

REFERENCES

Beamish, R. J., and D. R. Bouillon, Pacific salmon production trends in relation to climate, *Can. J. Fish. Aquat. Sci.*, *50*, 1002–1016, 1993.

Brodeur, R. D., and D. M. Ware, Long-term variability in zooplankton biomass in the subarctic Pacific Ocean, *Fish. Oceanogr.*, *1*, 32–38, 1992.

Confederated Tribes and Bands v. Baldridge, Civil Case C80-342, Order and Judgment of U.S. District Judge Barbara J. Rothstein, W.D. Wash., September 7, 1995.

Culbert, L., and J. Beatty, Salmon pact "disappointing," *Vancouver Sun*, final ed., June 4, 1999, p. A1.

Department of Fisheries and Oceans (DFO), Government of Canada, *Canada and U.S. Reach a Comprehensive Agreement under the Pacific Salmon Treaty*, press Release, and Backgrounders, available on-line, http://www.ncr.dfo.ca, June 3, 1999.

D'oro, R., Judge orders end to blockade Alaska officials attempt to end ferry standoff, *Anchorage Daily News*, final ed., July 21, 1997, p. A1.

Fraser River Sockeye Public Review Board, *Fraser River Sockeye 1994; Problems and Discrepancies*, Canada Communication Group Publishing, Ottawa, 1995.

Hardin, G., The tragedy of the commons, *Science*, *162*, 1243–1248, 1968.

Hare, S. R., and R. C. Francis, Climate change and salmon production in the Northeast Pacific Ocean, in R. J. Beamish (Ed.), *Climate Change and Northern Fish Populations*, Can. Spec. Publ. Fish. Aquat. Sci. 121, 1995; pp. 357–372.

Hare, S. R., N. J. Mantua, and R. C. Francis, Inverse production regimes: Alaska and West Coast Pacific salmon, *Fisheries*, *24*(1), 6–14, 1999.

Higgs, R., Legally-induced technical regress in the Washington salmon fishery, *Res. Econ. History*, *7*(1), 55–86, 1982.

Hogben, D., D. Rinehart, and J. Beatty, Prince Rupert fish boats end blockade of Alaskan ferry: Ringleaders comply with injunction after meeting late Monday evening with federal

fisheries minister David Anderson, and the ferry steams away, *Vancouver Sun*, final ed., July 22, 1997, p. A1.

Houghton, J. T., G. J. Jenkins, and J. J. Ephraums (Eds.), *Climate Change (1992)*, Cambridge University Press, Cambridge, 1992.

Kiladis, G. N., and H. F. Diaz. Global climatic anomalies associated with extremes in the Southern Oscillation, *J. Climate*, *2*, 1069–1090, 1989.

Latif, M., and T. P. Barnett, Decadal climate variability over the North Pacific and North America: Dynamics and predictability, *J. Climate*, *9*, 2407–2423, 1996.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, A Pacific interdecadal climate oscillation with impacts on salmon production, *Bull. Am. Meteorol. Soc.*, *78*(6), 1069–1079, 1997.

McKelvey, R., Game-theoretic insights into the international management of fisheries, *Natural Resour. Model.*, *10*(2), 129–171, 1997.

Miller K. A., Salmon stock variability and the political economy of the Pacific Salmon Treaty, *Contemp. Econ. Pol.*, *14*(3), 112–129, 1996.

Munro, G. R., T. McDorman, and R. McKelvey, Transboundary fishery resources and the Canada-United States Pacific Salmon Treaty, in *Occasional Papers: Canadian-American Public Policy*, Canadian-American Center, University of Maine, Orono, 1998.

Munro, G. R., and R. L. Stokes, The Canada-United States Pacific Salmon Treaty, in D. McRae and G. Munro, (Eds.), *Canadian Oceans Policy: National Strategies and the New Law of the Sea*, University of British Columbia Press, Vancouver, 1989, pp. 17–35.

O'Neil, P., Taking stock of the Salmon Treaty, *Vancouver Sun*, final ed., July 5, 1999a, p. A1.

O'Neil, P. Conservation card played key role in salmon deal, *Vancouver Sun*, final ed., June 5, 1999b, p. A1.Pacific Salmon Treaty, March 18, 1985, U.S.-Can 99 Stat. 7 [codified at 16 U.S.C. 3631–3644 (1997)].

Pacific Salmon Treaty, March 18, 1985, U.S.-Can., 99 Stat. 7 [codified at 16 U.S.C. 3631–3644 (1997)].

Pearcy, W. G., *Ocean Ecology of North Pacific Salmonids*, University of Washington Press, Seattle, 1992.

Polovina, J. J., G. T. Mitchum, and G. T. Evans, Decadal and basin-scale variation in mixed layer depth and the impact on biological production in the Central and North Pacific, 1960–88, *Deep Sea Res.*, *42*, 1701–1716, 1995.

Royce, W. F., An interpretation of salmon production trends in W.J. McNeil (Ed.), *Salmon Production, Management and Allocation*, Oregon State University Press, Corvallis, 1988.

Schmidt, Jr., R. J., International negotiations paralyzed by domestic politics: Two-level game theory and the problem of the Pacific Salmon Commission, *Environ. Law*, *26*, 95–139, 1996.

Trenberth, K. E., J. R. Christy, and J. W. Hurrell, Monitoring global monthly mean surface temperatures, *J. Climate*, *5*, 1406–1423, 1992.

Trenberth, K. E., and T. J. Hoar, The 1990–1995 El Niño–Southern Oscillation event: Longest on record, *Geophys. Res. Lett.*, *23*(1), 57–60, 1996.

Trenberth, K. E., and J. W. Hurrell, Decadal atmosphere-ocean variations in the Pacific, *Climate Dynam.*, *9*, 303–319, 1994.

U.S. Department of State, Diplomatic Note No. 0225 from Canada to the United States; reply; attached Agreement, available on-line, http://www.state.gov, June 30, 1999.

Yanagida, J. A., The Pacific Salmon Treaty, *Am. J. Int. Law*, *81*, 577–592, 1987.

Zhang, Y., J. M. Wallace, and N. Iwasaka, Is climate variability over the North Pacific a linear response to ENSO? *J. Climate*, *9*, 1468–1478, 1996.

CHAPTER 47

TRANSBOUNDARY RIVER FLOW CHANGES

ROGER S. PULWARTY

1 INTRODUCTION

Water is a "fugitive" resource in the sense that it flows naturally from one place to another, from one reserve to another (e.g., groundwater to surface), and from one physical state (solid, liquid, and gas) to another. Thus *transboundary* can mean many things including transitions from wet to arid zones, from upstream to downstream, from one country or province to the next, etc. The Convention on the Protection and Use of Transboundary Watercourses and International Lakes (1992) defines "transboundary waters" to mean "any surface or ground waters which mark, cross, or are located on the boundaries between two or more states." Wherever transboundary waters flow directly to the sea, these transboundary waters end at a straight line across their respective mouths between points on the low water line of their banks. Groundwater resources are also frequently shared by two or more countries. Emerging issues in water resources emanate from three categories of problems: (1) transboundary water availability, (2) transboundary groundwater allocation, management, and conservation, and (3) transboundary water quality (Caldwell, 1993). This chapter is primarily concerned with surface water in large river basins crossing international boundaries.

Watersheds can be defined by crossing a physical line (or bench), dividing surfaces from which waters flow in different directions to different outlets. The term '*river basin*' is here used to denote channel length and catchment area. '*Catchment area*' refers to a drainage area in which surface water flows to a common outlet channel. '*Watersheds*' can also be crossed conceptually, such as in differentiating between upstream and downstream emphases in managing water and sediment

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

flows, water quantity and quality relationships, changes from soil type to landscape type as a basis for study and administration, trade-offs between centralized versus decentralized emphases in management, and concern for impacts on the environment and on basins of origin (White, 1997).

Shared watersheds constitute about 47% of the global land area and are inhabited by about 40% of the world's population. Worldwide there are more than 214 major transboundary river basins, of which 74% are shared between only 2 riparian states (Kaufman et al., 1997). Thirteen basins are shared by 5 or more countries, 4 of these (the Congo, Danube, Nile, and Niger) are shared by 9 or more countries; 3 (the Zambezi, Amazon, and Rhine) are shared by 7 riparians; 4 (Chad, Volta, Ganges–Brahmaputra, Mekong) are shared by 6 countries, and 2 (La Plata, Elbe) are shared by 5 countries. Nine basins have 4 riparians each, and 30 major basins are shared by 3 countries. Of the 9 international water bodies shared by 6 or more countries, 5 are in Africa. The largest river basin in Europe, the Danube, is one-seventh the size of the Amazon Basin, but its waters are shared among 17 countries. On the smaller scale, there are over 400 surface waters between Finland and Russia alone, 150 crossing the Netherlands including the Rhine, and over 150 in the Ganges–Brahmaputra–Meghna system.

Much recent attention has been focused on highly visible water-related problems in the Middle East, the Indian Subcontinent, and the Aral Sea Basin. The UN Food and Agriculture Organization (FAO) has identified more than 3600 treaties relating to non-navigational water use dating between the years 1805 and 1984. Since 1945, approximately 300 treaties have been negotiated dealing with water management allocations of international basins (Kaufman et al., 1997). But, one-third of the major international rivers have no international agreement, and fewer than 30 have any cooperative institutional arrangements.

Rivers have constituted boundaries long before the rise of the modern nation state. Interestingly, the notion of "rival," derived from the Latin "rivalis," was originally used to describe people living on opposite banks of a river. Among the most important factors conditioning transboundary conflicts are the historical patterns of use and the needs of new "arrivals" and changes in values over time. Several countries now rely on significant amounts of surface waters originating outside of their national borders (Fig. 1). Changing social factors in particular have made water-related resource conflicts within countries common and, in situations where water is shared between two or more countries, increasingly unavoidable (Kaufmann et al., 1997).

Frederick (1996), Dinar (1997), and others have highlighted several factors underlying most international disputes involving river flows, including: the variability and uncertainty of supply, interdependencies among users, increasing water scarcity, overallocation and rising costs, the increasing vulnerability of water quality and aquatic ecosystems to human activities, ways and means of supplying safe water facilities, and the mobilization of financial resources for water development and management. Many of these issues derive from concerns in water resources management in general. How these concerns are met is strongly shaped by the choice of the spatial unit within which studies and management actions are conducted, by the way

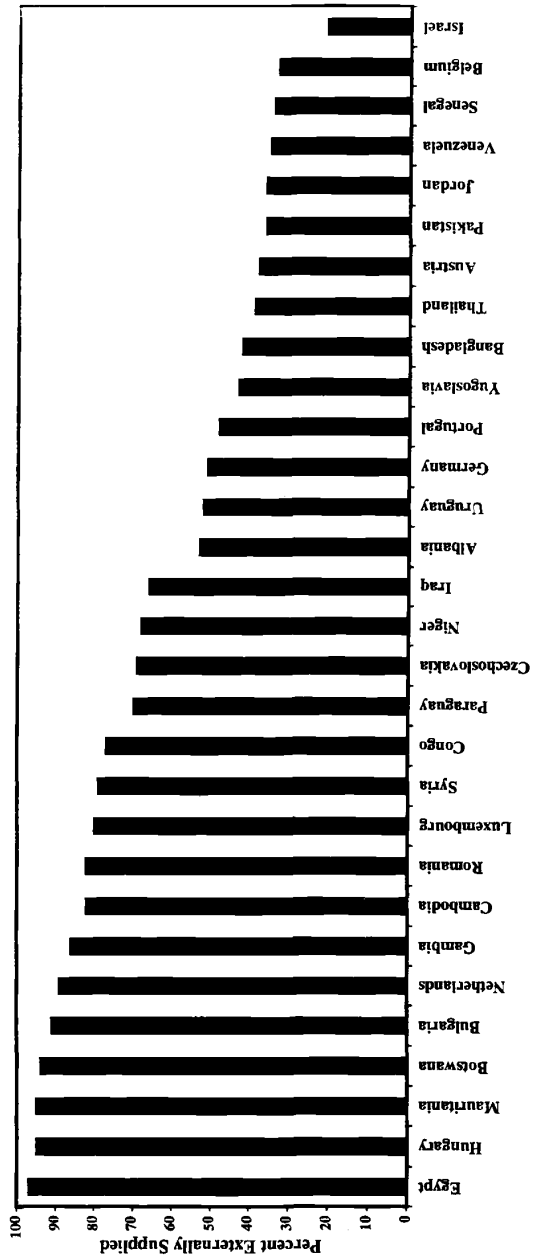


Figure 1 Percentage of water supply originating externally to selected countries (data from Gleick, 1993).

problems have been defined and changed over time, and by who benefits from defining problems in a particular way.

The primary concern in an international context is the need for international cooperation in the development of institutional and human resources for the efficient and equitable management of transboundary waters under variable and changing environments. In the following discussion the scales of human activities and interactions with large river basins are put in the context of streamflow changes on the time scales of century, decadal, seasonal, and extreme events.

2 IMPACTS

Transboundary fluctuations and changes in river flow can be attributed to (1) climate variations and change and (2) physical and biological transformations of basin hydrology including increased storage, diversions, and landscape changes. In this section, these conditioning factors on flow variability and change are discussed in general. Three cases are then selected for illustration in detail.

Climate Variability and Streamflow

At a given point along a river, streamflow is the product of the total catchment area above the gage and the average rate at which runoff is generated from snow and/or rain in that catchment. Runoff within a basin depends not only on rainfall but on its temporal distribution, vegetation cover (amounts and types), evapotranspiration, soil moisture storage capacity, rate of groundwater outflow, amount of paved area, etc. The seasonality of streamflow varies widely from river to river and is influenced mostly by the local seasonal cycle of precipitation, by timing of snowmelt (where appropriate), by the travel time of water from the runoff source areas, through surface and subsurface reservoirs and channels, and by large-scale human interventions.

Year-to-year variations in streamflow timing and magnitude play important roles in the development and management of water resources in most regions. Such interannual variations may be superimposed on longer (decadal to century-scale) fluctuations. Regime shifts from wetter to drier periods (and vice versa) on the decadal scale can be seen, for instance, in the annual streamflow data for the Colorado Basin in the southwestern United States (Fig. 2). Some but not all of these variations and their characteristics can be attributed to El Niño–Southern Oscillation events, climate forcings on the decadal time scale in the north Pacific, etc. There are many exceptions to large-scale patterns of streamflow seasonality in a region. Spatial differences in and between main streams and tributaries pose significant problems for flow estimation and planning. The shifts, and the surprises they introduce, provide a variable background against which allocations and water quality requirements are to be developed, agreed upon, and implemented.

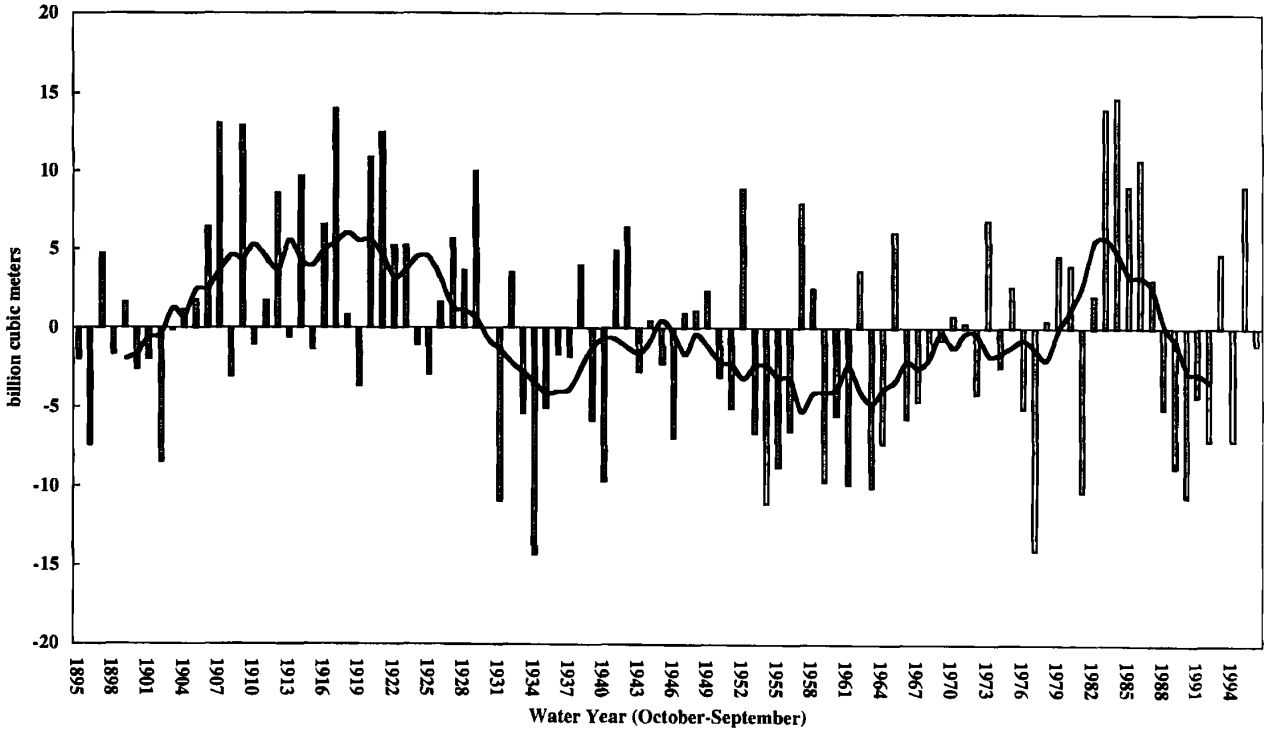


Figure 2 Colorado River streamflow: Annual deviations from the long-term mean (1896–1996) at Lee Ferry and 9-year moving average.

Physical and Biological Transformations: Storage, Diversions, and Landscape Changes

Human-built structures can either increase or decrease hydrologic connectivity of freshwater systems, rates of water movement, and transport and movement of organisms, materials, and heat. Construction of dams and diversions and modification of watersheds have greatly altered the natural flow regimes in streams and rivers (see, e.g., Fig. 3). Many existing flow regimes, particularly in large rivers, reflect human demands for water rather than natural cycles (Naiman et al., 1995). Two of the most dramatic changes to rivers in the twentieth century have taken place with regard to (1) water quality decline from return flows of agricultural and municipal waste water (L'vovich and White, 1990) and (2) large-scale diversions of water from one watershed to another. By 1990, the agricultural use of water worldwide was almost double that of all other uses combined, and canals in the former Soviet Union alone were estimated to be diverting more than $60\,km^3$ of water annually. In the United States diversions such as that from the Upper Colorado to Missouri and Rio Grande Basins and, more dramatically, from the Lower Basin to southern California have had significant impacts on the basins of origin and on streamflow into the Colorado River delta in Mexico.

At the global level, there has been a tripling of water use since 1950 (Postel, 1997). The number of large dams increased sevenfold to about 39,000, with reservoir capacity at about 9% of global annual river runoff. Until the 1930s, dam and reservoir designers were concerned primarily with single-purpose economic benefits (e.g., either transportation or irrigation). Since then, dams have been designed or altered to meet multiple-purpose criteria including flood control, hydropower, fishing, and recreation. Adverse social and environmental impacts include displacement, disease, siltation, scouring, reduced length of wild rivers, interference with migration and life cycles of aquatic species, introduction of exotic species, eutrophication and anoxia, and losses through evaporation and seepage.

Soil and vegetation act as intermediaries between precipitation and streamflow. Changes in landscapes brought about by urbanization, agriculture, forestry, industrialization, channelization, and construction of transportation corridors alter terrestrial and aquatic components of watersheds. Such alterations result in changes of flow (volume and timing) of water, sediments, nutrients and organisms in river channels, lake basins, wetlands, and groundwater. The rates, processes, and consequences of these changes are not well documented for most rivers. While the data on large-scale irrigated lands are for the most part reliable and comprehensive, no such detail exists for changes in drained wetlands and in low-lying grasslands.

In the following section, three cases (the Nile, the Colorado, and the Plata Basins) are chosen to illustrate problems, opportunities, and challenges in the sharing of transboundary streamflow under variable climate conditions and evolving management arrangements. By the late-1960s there were only 25 major rivers in the world with streamflow records extending back for at least 60 years (NAS, 1968). This discussion is limited to the reliable period of record for each basin considered.

3 THE NILE: CENTURIES OF CHANGE

The Middle East region, taken here to include the Tigris–Euphrates, the Jordan, and the Nile River Basins, has some of the highest population growth rates in the world and has a heavy reliance on irrigation for agricultural productivity. The region also has diverse historical religious, ideological, and ethnic disputes. Political friction and water scarcity have combined to produce perhaps the most volatile situation in the world (McCaffrey, 1993). At present, only the Tigris–Euphrates system has a significant amount of water left for allocation after present demands are met.

The Nile is the longest river in the world, flowing some 6500 km and draining an area of about 3.1 million km^2 spread over 10 countries. Most of its water is, however, generated over only 20% of this area. Its waters have been in use by Egypt for over 5500 years. The Nile differs from many transboundary rivers such as the Tigris–Euphrates and the Colorado River systems, in that the strongest economy, the strongest military force, and the best established water user in the basin, i.e., Egypt, is the downstream nation. Ninety-five percent of its runoff originates outside of Egypt (Table 1). The Nile experiences significant decadal to century-scale variability. Biblical stories of feasts and famines remind us of how the river dominated the climate experience and well-being of ancient Egypt (Riebsame et al., 1996).

The record of Nile flood levels at the Nilometer on Roda Island in Cairo has been called "the world's longest, best quantified climate proxy providing year-by-year values spanning an interval of 13 centuries" (see Diaz and Pulwarty, 1992). The Blue Nile flows from the Ethiopian Highlands to its confluence with the White Nile in Khartoum, Sudan. The White Nile loses half of its flow to the Sudan's Sudd (swamp lands region) as it leaves the Equatorial Plateau. The Atbara from Ethiopia is the last major tributary below Khartoum; here, the flow of the Nile averages 88 BCMY (billion cubic meters per year). From there the river flows 1200 km to Lake Nasser losing 4 BCMY to evaporation, with virtually no additional inflow for its remaining 2500 km run to the Mediterranean. Water levels are at their lowest from March to June and at their highest during the summer monsoon period from July to September. During August, the Ethiopian tributaries provide up to 95% of the flow. The Blue Nile as a watershed constitutes 16% of the entire basin, and at present contributes 62% of the annual flow at Cairo, while the White Nile contributes about 32%.

Changes in the contributions of the source regions to the aggregate streamflow have been noted during three main periods: (1) AD 650 to 1250 with about equal contributions from both sources, (2) AD 1250 to 1480 with greater contribution from the equatorial White Nile source, and (3) AD 1480 to 1870 with lower discharge from the equatorial source. Increases in flow from 1650 to 1750 were more widespread, roughly matching increases in Lake Chad levels. Variations also occur within these century-scale fluctuations on a 25- to 45-year time scale (Diaz and Pulwarty, 1992).

After the construction of the Aswan High Dam during the 1960s, the Nile maximum to minimum discharge ratio decreased from a natural condition of 12 : 1 to 2 : 1. In addition to reducing seasonal to interannual variability of the river and to providing hydropower and predictable flows for irrigation, the dam also reduced

sediment transport, estimated at 110 million tonnes per year, by 98%. The end result of these modifications in the basin is that the Nile delta is slowly declining and Egypt, at this time, uses more fertilizer per hectare than any other nation. In addition, 30 of the 47 commercial fish species available before construction of the Aswan Dam have been reduced to below harvestable levels.

Allocations of Nile water are based on the 1959 Agreement for the Full Utilization of the Nile Waters, which allotted 55.5 BCMY to Egypt out of a "fixed" flow of 84 BCMY. Requested allocations for upper Nile Basin states, most of which were British colonies in 1959, were rebuffed by Egypt. The result has been that Sudan and Egypt have harnessed most of the Nile's water with negligible development by other basin states. Egypt pursues a status quo position of fixed or increased inflows, resisting the construction of dams by Ethiopia on Nile tributaries. At present, Nile water in an average year plays a role in producing food for about half the population of Egypt. Large food imports, thus represent the importation of "virtual water," that is, water that Egypt would have had to obtain itself in order to grow the same amount of food.

As discussed by Riebsame et al. (1996), any calculation of climate change impact in the Nile Basin is complicated by assumptions about intricate water allocation and institutional arrangements, chiefly between Sudan and Egypt. At present, the Sudan requires additional Nile water for irrigation, but additional withdrawals would be detrimental to Egypt's already fully used supplies. Countries dependent on hydropower also have to agree to release adequate amounts of water for irrigation. However, there is considerable reluctance to release water at the expense of hydroelectricity. In addition, several of the upstream countries have severe food security problems, with much less money to import food than does Egypt (Appelgren and Klohn, 1997). These countries increasingly view water resources development, including interbasin transfer, as their major hope and option for socioeconomic development. One such system, the Jonglei Canal, designed to increase available Nile flows, is over half completed. The canal was designed to divert water around the Sudd to enable larger quantities of White Nile flow to reach Aswan. Its development is halted, at present. While rules for regulation of the equatorial lakes to minimize downstream and upstream losses have been developed, the scale of ecological impacts of the canal, if completed, are anticipated to be large for the Sudd. Even with better technical information, it is still up to the states of the Nile Basin to decide which operational strategy represents the most desirable compromise (Georgakakos et al., 1997). The Quaternary record also shows evidence that for periods of time the equatorial sources formed a closed basin, i.e., they did not flow to the Sudd nor reach the confluence. Pressures on Nile resources have now reached the stage where countries are obliged to make the river both an object and instrument of domestic and international politics (Appelgren and Klohn, 1997), without the flexibility needed to adapt to natural changes in variability or social trends. Ethiopia, for instance, is expected to have 25% more people than Egypt by 2025.

4 THE COLORADO: DECADAL-SCALE VARIATIONS

The past and present alterations of hydrology in the southwestern United States and northwestern Mexico reflect complex histories of human settlement, large-scale water diversions, the development and evolution of water policy and law, and expanding frameworks of water resources management. Rapid population increases, economic growth (including agriculture), the rise of urban centers over the last century (and more so recently) have resulted in intense pressures being placed on western lands, water, and institutions. In addition, the focus on water supply and the resulting design of water management agencies in the U.S. West evolved under the presumption of water as an open resource. Recent emphases on water demand management, on meeting obligations to Native American tribes, and on environmental concerns have altered the traditional roles of federal, state, and local agencies. Even in the water-rich Pacific Northwest region, trade-offs, for example (between hydropower and endangered species requirements), have brought allocation systems to their limits, threatening the very sense of community and reducing the likelihood of water transfers to drier regions.

The Colorado River flows from the high mountain regions of Colorado through Utah and Arizona to the Sea of Cortez in Mexico. Its major tributaries, the Green, flowing from Wyoming, and the San Juan give the Colorado a drainage area of about 629, 370 km^2. The Upper Basin (above Lee Ferry, Arizona), just below the state border with Utah, provides 83% of the annual flow. The Colorado does not discharge a large volume of water with estimated annual flows being about one tenth that of the Columbia and one-eighteenth that of the St. Lawrence in the northeastern United States, both of which drain basins of comparable size. It is, however, an important source of water for seven semiarid states in the western United States and for northwestern Mexico.

Significant human alteration to seasonal streamflow began with the development of the Yuma Valley Irrigation Project, which in 1912 provided irrigation water in Arizona and California. More importantly, the completion of the largest concrete dam in the world in 1935, the Hoover Dam on the Colorado, initiated the golden Age of Dams in the United States (Reisner, 1986). The Glen Canyon Dam (hereafter referred to as the GCD) was completed in 1963, physically dividing the Colorado River into its Upper and Lower Basins. The full reservoir system now stores about four times the annual average streamflow. The Upper Basin provides 80 to 90% of the total flow in the Colorado. The primary role of the GCD is to enable the Upper Basin states of Utah, Colorado, Wyoming, and New Mexico to utilize their apportionment of Colorado River water, while meeting obligations for water delivery to the Lower Basin states of Arizona, California, and Nevada, and also Mexico, consistent with the laws, treaties, compacts, and court decisions regarding Colorado River operations, collectively known as the Law of the River. Decadal-scale climatic factors influencing present water allocations are discussed in greater detail by Stockton and Jacoby (1978). Briefly, the period 1905 to 1930 was the wettest such period in 400 years of record, with 19.8 BCM constructed annual average flow at Lees Ferry. The Colorado River Compact (1922) among basin states used this average as

the base minimum for fixed allocation between Upper and Lower Basins. Since the signing of the Compact, the estimated annual virgin flow (1922 to 1997) has been 17.7 BCM, with an historic low flow of 6.9 BCM in 1934. During the 1931 to 1940 and 1954 to 1963 streamflow averaged about 12.6 BCM annually.

Under similar future conditions, if the Upper Colorado River Basin states consume the 9.3 BCM allocated to them by the Compact, they would default on the legal obligation to the Lower Basin. These Lower Basin states have the first right to the allotment. The engineering solution was to construct a dam (the GCD) near Lees Ferry that could store water in wet years and release water in dry ones.

Maintaining geopolitical equity between basins was therefore the major purpose served by the GCD (Ingram et al., 1990). Power generation itself was second to the need to generate revenue for other water projects primarily in the Upper Basin. Decisions in the Colorado River Basin now involve many temporal and spatial scales. A recent study by the consortium of western water resources institutes (i.e., the Powell Consortium) offers the counterintuitive result that while the Lower Colorado River Basin within the United States is indeed drier than the Upper Basin, it is the Upper Basin that is vulnerable to severe, long-term drought because of the 1922 agreement (Powell Consortium, 1995). In addition it was found that, while opportunities for "win–win" situations and rule changes exist, such changes are extremely difficult to implement. Minimum flow requirements are, at present, met with unused entitlements. As with many other river basins, in western North America, there is at present no single decision-making body that encompasses the entire basin.

Water managers have traditionally relied upon the historical record in order to plan for the future, inferring the probability that shortages and floods might occur given their frequency of occurrence in the past. As a result of the climatological droughts experienced during the 1930s, 1950s, and in 1977 (at 7.2 BCM the second driest year in the record), the system as a whole is operated to maximize the amount of water in storage for protection against dry years (see Fig. 3). Total storage within Lake Powell is over three times the annual Upper Basin allotment, or about 31.0 BCM. It should be noted that the region has also experienced sustained periods of high runoff as occurred from 1941 to 1950 and from 1983 to 1986.

The Colorado River, because of the scale of impoundments and withdrawals (including large-scale interbasin transfers), has been called the most legislated and managed river in the world. The river now virtually ends 16 to 30 km before reaching the Sea of Cortez. The impacts of these diversions and storage within Mexico are not well documented, but they have had the effect of disrupting fishing communities located along the river's delta and of decreasing water quality due to increased salinity. The 1944 Water Treaty between these two countries left important problems unresolved in the area of the quality of water delivered by the United States to Mexico. Indeed, the domestic realities at source regions, such as pollution and low flows, have forced the United States to assume costs for desalination of Colorado River water, before it enters Mexico. This has implications for possible long-term reductions in water availability during exceptionally dry periods. Since Mexico's entitlement to Colorado River water is less than 10% of the flow, it appears

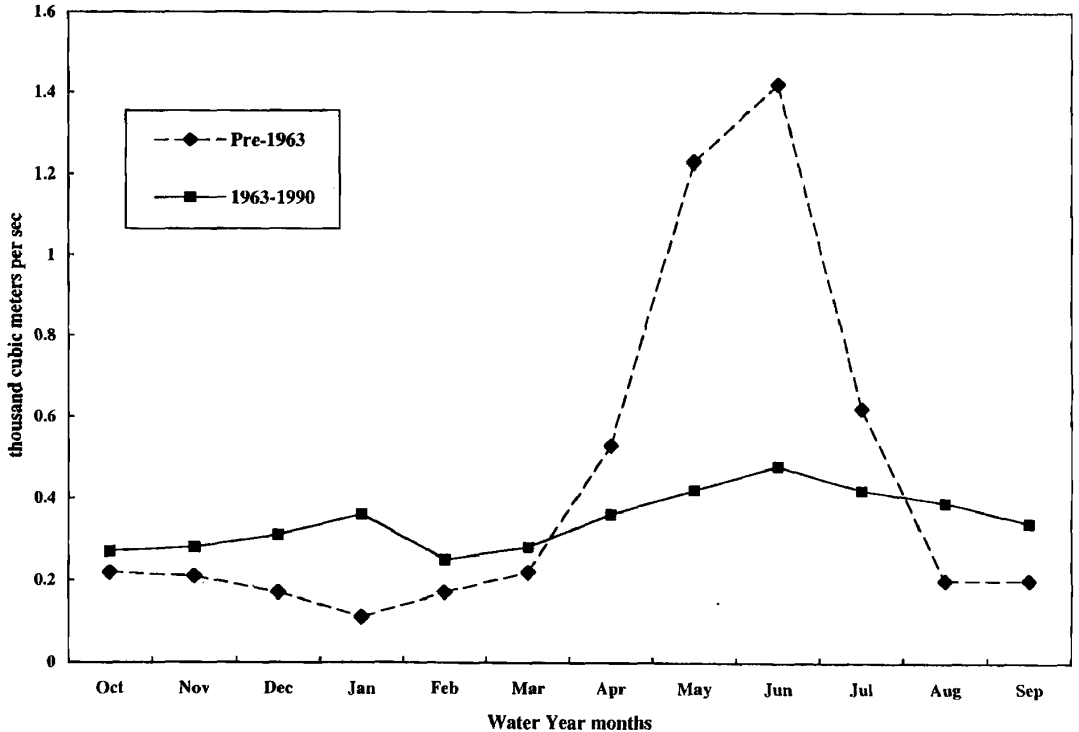


Figure 3 Colorado River discharge at Lee Ferry before and after completion of Glen Canyon Dam in 1963.

unreasonable to expect that Mexico will assume total responsibility for delta restoration. Such stresses are expected to produce serious conditions more often in the future because of the expansion of bilateral trade between the United States and Mexico, the projected rise of development and agriculture in western Sonora, and the full utilization of water allocations in the Upper Colorado Basin anticipated to occur by the year 2010.

5 THE PARANÁ–PARAGUAY RIVER BASIN: INTERANNUAL VARIABILITY AND EXTREME EVENTS

The Paraná–Paraguay River Basin in South America encompasses about 84% of the La Plata Basin with a population of about 100 million (1992 figures). It is the most developed agricultural and industrial zone in South America, accounting for 80% of the economic production in Argentina, Brazil, and Paraguay. The period of relatively small numbers of high floods between 1905 and 1960 coincided with the expansion of urban areas onto the valley bottoms. Three major changes have taken place in the basin since the 1960s (Penning-Rowsell, 1996). First, agricultural and industrial production has increased. The Upper Paraná Basin in Brazil has been converted from coffee plantations, developed in the 1940s, to fields of soybeans and sugar cane (for alcohol fuel production). Second, deforestation to establish cropland and pasture has been extensive in both Brazil and Paraguay. In Paraguay forested regions declined from about 45% in the eastern region in 1945 to about 15% in 1992. In northern Argentina and southern Brazil forest losses range from 60 to 90% in this century. Third, about 20 hydroelectric power plants have been built on the river, significantly changing system hydrology. For example, the width of the Paraná River varies dramatically from 4 km north of Guaira (Brazil) to about 60 m below the Itaipu Dam. As in other parts of the world, the increasing damages caused by extreme rainfall events result from urban and agricultural encroachment onto the floodplains, as much as from the events themselves. Most of these floodwaters (up to 85% on the Paraná) come from the Upper Basin in Brazil. Accompanying these changes are heavy sediment flows from agricultural lands bordering the Paraná.

The Upper Paraguay Basin has low river banks and is prone to flooding, creating a zone known as the 'Pantanal'. Flooding in the Paraná–Paraguay Basin has become both more frequent and more severe in recent years. In particular, the 1983 flooding cost an estimated US$1.8 billion, while 1992 flooding caused serious damage to infrastructure and capital stock resulting in estimated damages of up to US$1 billion and affecting 3.1 million people. The cost estimates of more recent and widespread flooding during the 1998 to 2000 period are as yet incomplete. Ten-year flood discharge rates are now over 15% greater than those in the early twentieth century (Anderson, 1993). In addition, the low-river flows in the watershed have been less frequent and less extreme (i.e., not as dry) in the latter part of the twentieth century. Five of the 10 largest floods recorded during this century (as measured by daily peak discharge) have occurred since 1982. At present El Niño (warm) events may provide the best explanation as a driver of increased precipitation during the rainy season

(March–May). The relative contribution of El Niño as opposed to La Niña–related teleconnections results in a difference of about 20% of annual streamflow. However, it is clear from the record that the general trend in streamflow is upward, and the hydrological regime of the rivers appears to have been changing primarily since 1940. It should be noted that there are also references to high flood events during the nineteenth century.

The human occupancy of these floodplains reflects their economic value for agriculture, communications, and transportation. Flood losses will increase because people are concentrated in flood-prone urban areas, there is increasing migration from rural areas, and land use is poorly regulated. Wet-season rainfall in the region has also been relatively high since the early 1980s. Good relations among Brazil, Paraguay, and Argentina are important for shared use of flood-forecasting data and mitigation strategies, including land-use changes and reforestation. Interagency rivalries within each of these countries restrict the capacity to act internationally, especially under crisis conditions. The prognosis for implementation of sustainable, preventive, nonstructural alternative approaches to reducing exposure (property at risk) and social vulnerability over the long-term does not appear promising.

6 PROBLEMS

Societies are always adapting incrementally and in diverse ways to a variety of integrated and cumulative changes. There is, however, little understanding of the long-term and widespread consequences of these adaptations (Dynesius and Nilsson, 1994). Questions remain as to how and when adaptation will occur and, in particular, how equity and environmental considerations will be addressed over time. A key component has been, and will continue to be, ensuring that the best available and most appropriate scientific information is employed in decision making (Pulwarty and Redmond, 1997). Unfortunately, detailed assessments of the direct human-induced changes of river hydrology of most large river systems are lacking, as are coherent assessments of discharge. Attempts to cope with the negative effects of technological interventions usually follow decades later (L'vovich and White, 1990).

Most major programs for complex utilization of rivers assume that average conditions for hydrologic records available at the time of project planning will continue indefinitely into the future (White, 1997). This has usually resulted in overestimations of supply or underestimations of demand over the long term. The complications of changes in the spatial and temporal distribution of rainfall, soil moisture, runoff, frequency, and magnitudes of droughts and floods have not been explicitly included in response planning. Systems design, operational inflexibility, and legal and institutional constraints reduce the adaptability of water systems to respond to climatic changes (Gleick, 1993). It is therefore difficult to plan for and justify expensive new projects on the basis of supply alone, when the magnitude, timing, and even direction of the changes in basins are unknown (Frederick, 1996).

The major stumbling blocks, exacerbated by shifts in climate or in the frequency of climate extremes, relate to the adversarial relationship that usually develops

between upstream and downstream users of water. The determination as to when a particular use is equitable and reasonable involves definitions of broad concepts such as "no harm" and "optimal utilization." Problems are further compounded by lack of agreement on event definitions, such as what constitutes an "extraordinary" (i.e., severe and persistent) drought in different places. The spatial extent and persistence of drought may produce shortages not only in the locale considered but also in neighboring regions that otherwise are supposed to make surplus water available through interbasin tranfers. These concepts appear clear from the standpoint of water measurements, but difficulties emerge in (a) the practical and equitable sharing of quality water or (b) how an upstream country should share water with downstream countries, especially during periods of water stress.

As pointed out by Frederick (1996) and others, in the absence of clear and enforceable property rights, the strongest, most clever, and most advantageously positioned countries can claim and use scarce resources with little concern for the impacts on others. Scarcity does not arise on its own but is dependent on the quantity and quality of resources at particular times and on the degree of access to and capacity to use those resources. The call to the market as a first order of business does not address historical inequities, political differences, and environmental needs. The most powerful usually have the greatest resources to purchase property rights anyway.

All of the river basins described above exhibit characteristics of "closed or closing" water systems. In such systems the management of interdependence becomes a public function; development of mechanisms to get resource users to acknowledge interdependence and to engage in negotiations and binding agreements become necessary. The implementation of such mechanisms does not appear to be viable without focusing events (Keller et al., 1992). Focusing events, such as the La Plata floods, are usually associated with exceptional societal or environmental impacts, highlight critically vulnerable conditions, and elicit highly visible responses. Clusters of historic events may combine with physical events to precipitate or allow particular actions to be undertaken e.g. the cumulative roles of high flows prior to 1930, the 1930s drought, the Great Depression, World War II, and the 1950s drought in influencing dam building and management on the Colorado). Experience shows that decisions bringing rigidity to the management system ultimately generate more problems than they resolve (see in addition to cases cited here, Glantz, 1988; Gunderson et al., 1994). As is evident on the Lower Colorado River Basin and the Nile, early "winners" are unlikely to be willing to alter earlier terms of agreement even when changes in climate conditions are well documented. These problems are further complicated by the unique context imposed by transboundary resources at the borders themselves. According to Ingram et al. (1997), political boundaries, whether domestic or international (1) often separate the location where problems are felt from the location where the most effective and efficient solutions can be applied; (2) make restraint in using scarce water resources unlikely especially when the forces of global economic competition reinforce the focus on opportunities for immediate economic profit; (3) aggravate perceived inequalities; and (4) obstruct grassroots

problem solving. In addition, they note that national and state policies are usually at odds with border needs and priorities.

As pointed out by one geographer, "it would be naive to envisage government-sponsored research on the cultural conflicts and politics of water management" (Wescoat, 1991 p. 392). The end result is that in the absence of explicit discussion of conflict the policy process is pushed toward the "technological fix," and time and resources are allocated to achieve near-term tangible results rather than long-term solutions (Caldwell, 1993). Practice thus becomes largely issue specific and incremental with the focus on winners and losers rather than on the development of a consensual vision of a preferred future. There is, in addition, limited experience for managing impacts of severe events, such as persistent drought, in the context of projected rates of development or in the context of closed international water systems. In meeting new challenges and trade-offs brought by a changing and variable environment and societal changes, we simply do not know how (precisely) we must plan.

7 LESSONS

As discussed above, climate and weather events form a variable background on which agreements, conflicts, and politics are constantly being played out. Demographic, political, and environmental changes can and do disrupt existing relationships and current wisdom about the interactions between society and the environment (Glantz, 1994). There have, however, been cases where regional cooperation has led to particular solutions. In 1996 the UN Economic Commission Convention on the Protection and Use of Transboundary Watercourse and International Lakes came into being (Wieriks and Leidig, 1997). Parties to the Convention are obligated to prevent, control, and reduce water pollution, primarily the influx of hazardous materials from point and nonpoint sources.

Lessons provided by the Convention and from earlier treaties (e.g., the 1960 Indus Treaty between Pakistan and India) lead to the following conclusions: (1) international water problems can only be effectively handled on the river basin scale with full acknowledgement of interdependence, (2) river basin management requires an overall integrated approach, including attention to ecological water quality and water quantity issues, (3) international strategies and policies should leave room for flexible implementation, (4) public and political support are also prerequisites for successful formulation, particularly regarding environmental policies, (5) major decisions cannot be taken without input from all stakeholders and ensuring adequate legal basis for participation, and (6) cooperation will not occur without mutual confidence among all parties involved. Most importantly, implementation must be explicitly provided for, and usually does not succeed without some shared vision for the future.

Recognition of variability and change in water resources is a first step in accepting that management occurs under changing conditions in which surprise and uncertainty will always exist. From the brief reviews provided above, several conclusions

can be drawn about climate-weather and water relationships: (1) it is unwise to ignore the variability that is inherent in natural systems, since decisions that bring rigidity to a management system can ultimately generate more problems than they resolve; (2) it is important not to ignore changes that have and will occur in social systems; (3) major changes in streamflow can be regarded in retrospect as climate changes, and; (4) careful examination of past seasonal to decadal-scale variability and responses can provide useful organizational lessons for areas with increased or decreased water supply, as may be postulated under different climate change scenarios. Expectations about the future tend to be better understood by people within organizations if there is a clear parallel with the past. Incentives to conserve and opportunities to reallocate supplies as conditions change do not require long lead times, large financial commitments, or accurate information about the future (Frederick, 1996). There is, however, a clear need for exchange of experience and learning among different basins especially on how awareness of slow onset problems in the context of decadal-scale variability is developed and the ways in which societies have adjusted to them. A conspicuous aspect of water management has thus been the lack of careful postaudits (systematic and iterative evaluations) of the social, economic, and consequences of previous programs and ongoing projects (White, 1997).

Few assessments, intended to provide insight into future responses, show sensitivity to historical dimensions. For instance, it is impossible to understand the present context without acknowledging that for most of the twentieth century, following the disintegration of the Ottoman Empire and the post–World War I period, water disputes in the Middle East were closely associated with boundary drawing, state formation, nation building, domestic and international strife, and security issues. Attention to history shows that people have known about most problems for a long time but have not acted on better knowledge of these past changes, i.e., problems have been accumulating for many years not just when they are publicized (Glantz, 1999). Inattention to these changes or engaging in inadequate responses allows the incremental accumulation of problems to the point of system criticality or collapse. Mitigating future impacts requires greater emphases on social and ecological factors that prefigure these "surprises."

As a cautionary note, the idea that many solutions to reducing social and environmental vulnerabilities, cognizant of physical, social, and economic time frames, are available but remain unused is not new (Ascher and Healy, 1990; Pulwarty and Riebsame, 1997). There are no apparent quick fixes, technological, economic, or otherwise (Glantz, 1999). A better understanding of the links between domestic political concerns and foreign policy is needed in order to construct a more complete picture of issues underlying water disputes. One of the most important benefits that may be realized through a comparative study is an understanding of why some policies may be chosen over others, how these are related to particular climatic events, which ones rise to prominence, and which are allowed to persist. Identification of the barriers to implementation and evaluation in one setting may shed light on the likelihood of success of similar actions in another setting.

8 IMPORTANCE OF LINKING HUMAN AND PHYSICAL ASPECTS

Cultural, political, and economic conflicts are the most serious problems encountered in transboundary river issues. These may be exacerbated by environmental stresses and by the actions chosen in response. Gilbert White (1966) distinguished between the theoretical and practical ranges of choice in structuring the analysis of adjustment decisions. The physical environment at a given stage of technology sets the theoretical range of choice open to any resource manager or group. The practical range of choice is set by culture, institutions, types of analytical tools employed, etc., which permit, prohibit, or discourage a given choice. Water has thus been defined as a property of territorial units in the legal setting, as a natural resource transformable into products for human consumption in an engineering setting, and as a commodity that can be exchanged and traded between various places and various uses in an economic setting (Blatter and Ingram, 2001). The technological response is to regularize the flow and expand the total amount of available water. The recent experience on dispute resolution shows a contrast between an abiding belief in the rationality of public decision making on the part of some participants while others exhibit sharp suspicion toward the politics of "expert" managerial discourses (e.g., the commodification of water vs. communal values). Emerging knowledge of the complexity and varieties of meaning characterizing "fresh water" at the end of the twentieth century has led to greater appreciation of the limitations as well as the benefits of technological, legal, and economic approaches. In no other context are the divergent meanings of water likely to be more contested than in transboundary situations (Blatter and Ingram, 2001).

Attention has begun to focus on how impacts of chosen approaches exacerbate the root causes of social and ecological vulnerability. There is increasing appreciation for explicit consideration of peoples affected by decisions but who are usually excluded from participation or from the benefits of developed infrastructure (see Pulwarty and Riebsame, 1997; Milich and Varady, 1998). Reduction of vulnerability requires careful assessments of the range of alternative adjustments, among which societies may choose in arriving at a suitable plan for a given period (White, 1977). There is thus an ongoing need for guidance in integrating equity with efficiency considerations in transboundary water management through studies of place-based historical and cultural uses, understanding the role of public trust, the impact of new technologies, and of the flexibility provided by market-based approaches after basic human needs and environmental requirements are met. In particular more work needs to be done on the trade-offs involved between calls for increased participation and the formation of consensus. Promising partnerships are emerging but are in their early stages (see Milich and Varady, 1998).

From the perspective of the climate and water sciences, researchers, through ongoing dialog and joint studies, should engage practitioners as full partners to uncover issues of mutual significance, explicitly address uncertainties in both the scientific and decision domains, and to understand and overcome barriers to information use contingent in each situation (Pulwarty and Melis, 2001). The goals are to have better matches among what is needed, what is asked for, what is available, and

what actions can be taken. These processes must be embedded within an understanding of the decision contexts (historical, policy, and operational) within which trade-offs take place.

Water by its very nature tends to introduce even hostile co-riparians to cooperate even as disputes continue over other issues (Wolf, 1999). At the international level the weight of historic evidence tends to favor water as a catalyst for cooperation for particular ends. This has not been the case on the subnational scales. While governing institutions that more closely correspond with the physical water system can help to assure appropriate consideration of efficiency and equity, domestic policy can pose major institutional barriers to international agreements and management across national borders. Ultimately, the main tasks in the foreseeable future will be uncovering how to share common but variable water resources in a catchment area between upstream and downstream users, between various sectors, between rural and urban areas, between preservation of functioning ecosystems, and more direct tangible needs (Falkenmark and Lundqvist, 1995). Engaging the many dimensions of transboundary river flow requires, more than ever, the need to understand these "regions" as integrators of social, cultural, climatic, economic, and ecological histories and networks, that help to shape shared community interests and values.

REFERENCES

Anderson, R., *An Analysis of Flooding in the Paraná/Paraguay River Basin World Bank*, LATEN #5, World Bank, Washington, DC, 1993.

Appelgren, B., and W. Klohn, Management of transboundary water resources for water security: Principle, approaches and state practices, *Nat. Resour. Forum*, *21*, 91–100, 1997.

Ascher, W., and R. Healy, *Natural Resource Policymaking in Developing Countries*, Duke University Press, Chapel Hill, NC, 1990.

Blatter, J., and H. Ingram (Eds.), *Reflections on Water: New Approaches to Transboundary Conflicts and Cooperation*, MIT Press, Cambridge, 2001.

Caldwell, L., Emerging boundary environmental challenges and institutional issues: Canada and the United States, *Nat. Resour. J.*, *33*, 10–31, 1993.

Correia, F., and J. da Silva, International framework for the management of transboundary water resources, *Water Int.*, *24*, 86–94, 1999.

Diaz, H., and R. Pulwarty, A comparison of Southern Oscillation and El Niño signals in the tropics, in H. Diaz and V. Markgraf (Eds.), *El Niño: Historical and Paleoclimate Aspects of the Southern Oscillation*, Cambridge University Press, 1992, pp. 175–192.

Dinar, A., Economic and social considerations of regional cooperation in River Basin comprehensive water resources development, keynote presented at Nile 2002 Conference, Addis Ababa, Ethiopia, February 24–28, 1997.

Dynesius, M., and C. Nilsson, Fragmentation and flow regulation of river systems in the northern third of the world, *Science*, *266*, 753–762, 1994.

Falkenmark, M., and J. Lundqvist, Looming water crisis: New approaches are inevitable, in L. Ohlsson (Ed.), *Hydropolitics: Conflicts over Water as a Development Constraint*, Zed Books, London, 1995.

Frederick, K., Water as a source of international conflict, *Resources*, *123*, 9–19, 1996.

Georgakakos, A., W. Klohn, and K. Georgakakos, A decision support system for the Nile River, in *Proc. 5Th Nile 2002 Conference*, Addis Ababa, Ethiopia, February 24–28, 1997.

Glantz, M., H., *Creeping Environmental Problems and Sustainable Development in the Aral Sea Basin*, Cambridge University Press, Cambridge, 1999.

Glantz, M. H. (Ed.), *The Role of Regional Organizations in the Context of Climate Change*, Springer-Verlag, Amsterdam, 1994.

Glantz, M., *Societal Responses to Regional Climatic Change: Forecasting by Analogy*. Westview Press Colorado, 1988.

Gleick, P. (Ed), *Water in Crisis*, Oxford University Press, New York, 1993.

Gunderson, L., C. Holling, and S. Light, *Barriers and Bridges to the Renewal of Ecosystems and Institutions*. Columbia University Press, 1994, pp. 3–34.

Ingram, H., L. Milich, and R. Varady, Managing transboundary resources: Lessons from Ambos Nogales, *Environment*, *36*, 6–38, 1994.

Ingram, H., D. Tarlock, and C. Oggins, The law and politics of the operation of Glen Canyon Dam, in *Colorado River Ecology and Dam Management*, National Research Council, National Academy Press, 1990, pp. 10–27.

Kaufman, E., J. Oppenheimer, A. Wolf, and A. Dinar, Transboundary fresh water disputes and conflict resolution: Planning an integrated approach, *Water Int.*, *22*, 37–48, 1997.

Keller, J., N. Peabody, D. Seckler, and D. Wichelns, *Water Policy Innovations in California*. Center for Economic Policy studies, Winrock International 1992.

L'vovich, M., and G. White, Use and transformation of water systems, in Turner et al. (Eds.), *The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the Past 300 Years*, Cambridge University Press, 1990; pp. 235–252.

McCaffrey, S., Water, politics and international law, in P. Gleick (Ed.), *Water in Crisis*, Oxford University Press, New York, 1993, pp. 92–104.

Milich, L., and R. Varady, Managing transboundary resources: Lessons from transboundary accords, *Environment*, *40*, 10–41, 1998.

Naiman, R., J. Magnuson, D. McKnight, and J. Stanford, *The Freshwater Imperative: A Research Agenda*, Island Press, Washington, DC, 1995.

National Academy of Science (NAS), *Alternatives in Water Management*, National Academy Press, Washington, DC, 1968.

Penning-Rowsell, E., Flood-hazard response in Argentina, *Geogr. Rev.*, *86*, 72–90, 1996.

Postel, S., *Last Oasis: Facing Water Scarcity*, W. W. Norton, New York, 1997.

Powell Consortium, Severe sustained drought: Managing the Colorado River in times of water shortages, *Water Resour. Bull. Spec. Issue*, *31*(5), 1995.

Pulwarty, R., and T. Melis, Climate extremes and adaptive management on the Colorado River. *J. Environ. Mgmt.*, *63*, 307–324, 2001.

Pulwarty, R., and K. Redmond, Climate and salmon restoration in the Columbia River basin: The role and usability of seasonal forecasts, *Bull. Am. Meteorol. Soc.*, *78*, 381–397, 1997.

Pulwarty, R., and W. Riebsame, The political ecology of natural hazards, in H. Diaz and R. Pulwarty (Eds.), *Hurricanes: Climate and Socio-Economic Impacts*, Springer-Verlag, 1997.

Reisner, M., *Cadillac Desert*, Penguin Books, 1986.

Riebsame, W. E., K. M. Strzepek, J. L. Wescoat, Jr., R. Perritt, G. L. Gaile, J. Jacobs, R. Leichenko, C. Magadza, H. Phein, B. J. Urbiztondo, P. Restrepo, W. R. Rose, M. Saleh, L. H. Ti, C. Tucci, and D. Yates, Complex river basins, in K. Strzepek and J. Smith (Eds.), *As Climate Changes: International Impacts and Applications*, Cambridge University Press, 1996, pp. 57–91.

Wescoat, J. L., Managing the Indus Basin in light of climate change, *Global Environ. Change*, *1*, 381–395, 1991.

White, G. F., Formation and role of public attitudes, in M. Jarrett (Ed.), *Environmental Quality in a Growing Environment*, Johns Hopkins Press, Baltimore, MD, 1966, pp. 105–127.

White, G. F., *Environmental Effects of Complex River Development*, Westview, Boulder, CO, 1977, pp. 1–22.

White, G. F., Watersheds and streams of thought, in H. Barakat and A. Hegazy (Eds.), *Reviews in Ecology: Desert Conservation and Development*, Cairo, Egypt, 1997, pp. 89–98.

Wieriks, K., and A. Schulte-Wulwer-Leidig, Integrated water management for the Rhine, *Nat. Resour. Forum*, *21*, 155–156, 1997.

Wolf, A., Conflict and Cooperation along International Waterways, *Water Policy*, 1&2, 251–265, 1998.

CHAPTER 48

LESSONS FROM THE RISING CASPIAN

IGOR S. ZONN

1 INTRODUCTION

The Caspian Sea is the biggest inland body of water in the world. Its surface area is roughly equivalent to the combined area of the Netherlands and Germany (about 400,000 km^2, or 144,000 mi^2). The surface water inflow into the sea is formed by the flow of the Volga, Ural, Terek, Sulak, Samur, Kura, small Caucasian rivers, and Iranian rivers. The watershed area of the Caspian Sea is 3.5 million square kilometers. The basin of the Volga River makes up nearly 40% of the territory of the catchment of the Caspian Sea, and it supplies about 80% of the total volume of annual water flow into the sea. All components of the Caspian ecosystem, directly or indirectly, to a greater or lesser extent, are influenced by river flow.

The Caspian Sea basin falls into three morphologically different parts: (I) the northern (25% of the sea area), a shallow area (less than 10 m deep; about 20% with depths less than 1 m) extending to a conventional line passing from the Terek river to the Mangyshlak Peninsula; (II) the medium (35%), with an average depth of 170 m (the maximum being 790 m); and (III) the southern (39%), the deepest area, with a maximum depth of 1025 m and an average depth of 325 m (see Fig. 1). Deep depressions in the northern and southern parts of the sea are divided by an underwater threshold running from the Apsheron Peninsula to Turkmenbashi (formerly Krasnovodsk) (Kosarev and Yablonskaya, 1994).

Before the breakup of the Soviet Union in December 1991, the USSR and Iran were the only two independent nations occupying the shores of the Caspian. With the breakup, three additional newly independent nations emerged along the coast: Azerbaijan, Kazakstan, and Turkmenistan. The Russian Federation's Caspian coastline is shared by three of its political units: Astrakhan Oblast, the Republic of Kalmykia, and the Republic of Dagestan.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

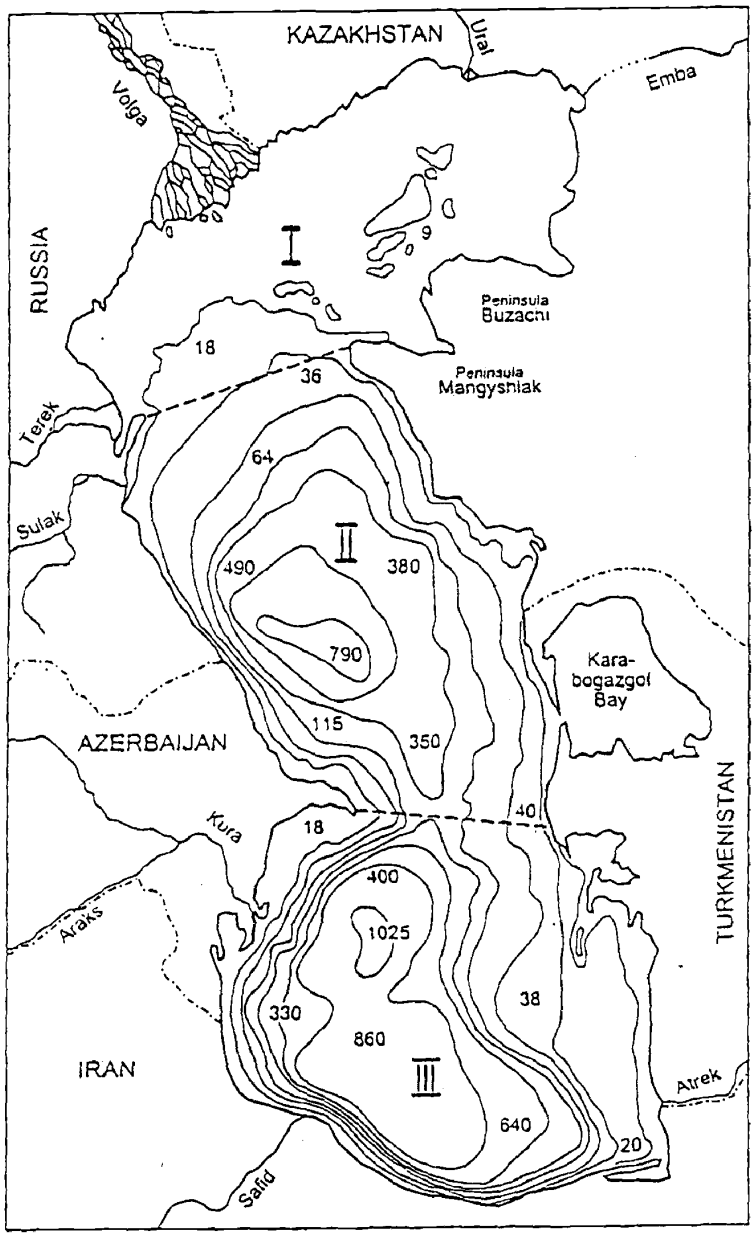


Figure 1 Depth isolines for the Caspian Sea, in meters.

2 NATURE OF SEA-LEVEL CHANGES IN CASPIAN SEA

The Caspian Sea is a closed basin in the inland part of Eurasia and this sea's water level is below that of the world ocean. The sea basin stretches almost 1200 km from north to south and its width varies between 200 and 450 km. The total length of the coastline is about 7000 km. Its water surface area is about 390,600 km^2 (as of January 1993). Water salinity in the northern part is 3 to 6‰, and reaches 12‰ in the middle and southern parts.

Fluctuations in sea level for various lengths of time can be found in the data of geomorphological and historical studies of the record of the Caspian Sea (Fig. 2). Within the last 10,000 years, the amplitude of fluctuations of Caspian Sea level has been 15 m (varying from −20 to −35 m). During the period of instrumental observations (from 1830 onward), this value was only about 4 m, varying from −25.3 m during the 1880s to −29 m in 1977. Annual increases in the level during this period met or exceeded 30 cm on three occasions (in 1867, 30 cm; in 1979, 32 cm; and in 1991, 39 cm). The mean annual increment in the level in the 1978 to 1991 period was 14.3 cm.

Natural factors are the primary cause of recent Caspian Sea level fluctuations (but not the only cause). Scientists have identified three distinct periods of level changes: 1830 to 1930, 1931 to 1977, and 1978 to the present. The first period of 100 years saw sea-level fluctuations not exceeding 1.5 m (5 ft). Researchers considered this period to have been relatively stable. The second period, from 1931 to 1977, is identified by a constant decline in level by 2.8 m (9.1 ft), and in 1977 the Caspian Sea reached its lowest level since the beginning of instrumental record-keeping in the 1830s.

As the sea level declined throughout the 1950s, 1960s, and early 1970s, Soviet scientists forecast that the decline would continue for at least a few decades into the future. Scientists have linked the reason for the decline to the regulation of Volga River flow. During these decades, major engineering activities were undertaken along the Volga, such as the construction of water diversion canals, reservoirs,

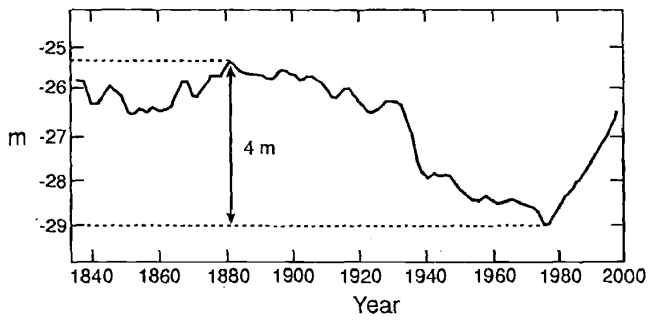


Figure 2 Caspian Sea level, 1835–1999, observed.

and dams. The construction of such engineering facilities diverted water away from the Caspian.

In response to this major drop in sea level, human settlements bordering the sea coast began to move toward the receding coastline. Fields and pasturelands were prepared for use, roads and rail lines were constructed, and housing and factories were built on the newly exposed seabed. During the Soviet era, many people emigrated from other parts of the region to settle along the border of the sea. Development of infrastructure along the coast took place to support the increasing population.

In an attempt to save the Caspian from drying out, Soviet scientists and engineers proposed the construction of a dam to block the flow of Caspian water to Kara-Bogaz-Gol Bay, a large desert depression in Turkmenistan adjacent to the sea's eastern shore. Political decisions made in the mid-1970s ordered the construction of the dam, but due primarily to bureaucratic inertia, the dam was not completed until the early 1980s. This was a few years after the Caspian's sea-level change had reversed direction. Before the dam was constructed, the bay took in 40 km^3 (8.6 mi^3) of Caspian water annually. It served as a huge evaporation pond, as well as a natural location for the accumulation of commercially useful mineral salts.

Another Soviet government response to the decline in the Caspian's sea level was a diversion of water into the Volga River from other Soviet rivers that flowed northward into the Arctic Ocean. River water flowing into the Arctic was viewed as wasted and without value to the Soviet Union because it was unused by human activity.

3 THE CASPIAN RISES

To the surprise of Soviet scientists, the level of the sea began to rise suddenly in 1978, the beginning of its third period of level changes. Since then, the Caspian has risen steadily by more than 2.5 m. One of the first actions the newly independent government of Turkmenistan took in 1992 was to tear down the dam in order to allow great amounts of water to flow into Kara-Bogaz-Gol Bay again and to replenish the supply of salts.

Scientists have proposed a variety of hypotheses about why the Caspian Sea level had increased so rapidly. These can be clustered into the following categories: tectonic plate movement on the seabed, climate fluctuations and change, and hydraulic construction along the Volga River, or some combination of these factors.

Tectonic Plate Movement Hypothesis

Tectonic movements over periods such as centuries and millennia have been the cause of many geologic changes in the Caspian basin. The region has been subjected to uplift, subsidence, overthrust of landforms, seabed mud-volcanic activity, and landslides, in addition to erosion processes and the accumulation on the Caspian seabed of river-transported sediments. However, it is difficult to see how tectonic

movements could cause such sharp fluctuations in the Caspian's sea level over relatively short periods. Thus, it appears that such movements have had an insignificant impact on recent sea-level fluctuations.

Climate Change Hypothesis

Today, most Russian scientists believe that climatic factors are the real cause of the Caspian Sea level rise. Studies by Golitsyn (1989) and Golitsyn and McBean (1992) indicate that recent changes of the Caspian Sea level are 90% associated with corresponding changes in the water balance components of the sea, as opposed to possible tectonic activity. The volume of inflow from rivers to the sea increased sharply after 1978. During certain years (e.g., 1979, 1985, and 1990), more than $350\,km^3$ of river water entered the sea. From 1978 until 1990, Volga River flow exceeded $260\,km^3/yr$. At present, no arguments have challenged the view that the main contribution to seasonal and annual level fluctuations of the Caspian is accounted for by surface inflow and evaporation levels. Within recent decades, the sea's fluctuations have been subjected to anthropogenic impacts as well.

In this regard, climate has two dimensions: climate fluctuations and climate change. Climate fluctuations occur on various time scales, with those of interest to present-day society being on the order of decades and perhaps centuries. Climate-related fluctuation refers to the increase and decrease of sea level over the course of decades. During the past two centuries, the sea has undergone several fluctuations. Those of the twentieth century have adversely affected socioeconomic activities and infrastructure along the sea's coastline.

The view that climatic processes in the Volga basin are the dominant cause of sea-level fluctuations has been recently reinforced. Droughts in this basin and sharply reduced Volga flow into the Caspian from mid-1995 until early 1997 have been associated with a 25-cm (10-inch) drop in Caspian level. Nevertheless, Russian scientists still suggest that the sea level will continue to rise into the first decades of the twenty-first century.

Climate change associated with global warming induced by human activities has also been proposed as the forcing factor behind the Caspian's rise since 1978. Those who see global warming as the forcing factor suggest that the most recent sea-level rise can be associated with intensification of the hydrologic cycle (i.e., more active precipitation-producing processes), an intensification that some scientists have linked to the human-induced global warming of the atmosphere. An increase in precipitation within the Volga River basin would translate into increased sea level.

Hydraulic Construction Hypothesis

Some observers have argued that the recent fall and rise in sea level were the result of human activities. They suggest that the widespread development of hydraulic structures (e.g., dams, reservoirs, irrigation systems) in the Volga River basin, beginning in the 1950s, led to a sharp decline in Volga flow. The filling of many reservoirs built along the rivers flowing into the Caspian, the increase in industrial and muni-

cipal water use by several times, and changes in the water regime of the floodplains led to a decrease of streamflow into the sea. Such a hypothesis could be tested by constructing a water budget model for the Caspian. Such a model would need to identify all the inflows into the Caspian Sea (such as from rivers and groundwater) and all outflow from the sea (such as evaporation and water diversions). While it is a seemingly straightforward task, identifying all the sources and sinks of Caspian water is not easy.

There is also a hypothesis about an Aral Sea connection. Yet another suggestion that seems to be made at just about every Aral or Caspian Sea conference is that the decline in the level of the Aral Sea is linked to the rise in level of the Caspian. The reasoning is that water diverted from the Aral basin to the Caspian basin to irrigate the desert sands for cotton production in Turkmenistan ends up either being evaporated into the air or seeping into the groundwater, which eventually makes its way into the Caspian. However, it is important to point out that *both* the recent fall and rise in the Caspian Sea level occurred during three and a half decades of a constant decline in the Aral's level.

4 SOCIETAL IMPACTS OF SEA-LEVEL RISE

According to a UN Environment Programme estimate, the cost of the impact of the sea-level rise of the Caspian, as of 1994, was $30 to $50 billion (US). Coastal ecosystems have been destroyed, villages inundated and populations evacuated, sea banks eroded, and buildings destroyed. Coastal plains have been invaded by subsurface seawater or have become waterlogged. Fauna have changed, and pasturelands and sturgeon spawning grounds have been destroyed.

Each of the five countries sharing the coasts of the Caspian Sea has suffered losses, and those losses increased until the mid-1990s. They suffer from the different impacts of sea-level rise because the territory along its coastline is neither uniformly settled nor uniformly developed economically. Economic losses in the big cities and villages have been higher than in the rural areas. More specifically, in Astrakhan Oblast (equivalent to an American state), about 10% of its agricultural land was out of production by 1995 because of sea-level rise. The coastline of the Republic of Dagestan (also part of Russia) was affected by the flooding of at least 40 factories in its cities of Makhachkala, Kaspiysk, Derbent, and Sulak. Nearly 150,000 hectares (370,000 acres) of land have been inundated, with a loss of livestock production and breeding facilities. Much of the 650-km (390-mile) Caspian coastline of Turkmenistan is made up of low-lying sandy beaches and dunes that are vulnerable to coastal flooding and erosion. In fact, some Turkmen villages that were once several kilometers from the sea are now coastal communities. Similar adverse impacts of sea-level rise on human settlements and ecosystems are found in Kazakstan, Azerbaijan, and Iran.

The Caspian has been referred to as a "hard currency sea" because of its large oil and natural gas reserves and because of its highly valued caviar-producing sturgeon. Regional reserves contain upward of 18 billion metric tons of oil and 6 billion cubic

meters (215 billion cubic feet) of natural gas. Experts suggest that the Caspian is second only to the Persian Gulf with respect to the size of its oil and gas reserves, and that Turkmenistan is a "second Kuwait." If the sea level were to continue to rise, a large part of the oil and gas mains along the Turkmen coast would become submerged and would also be subjected to corrosion by seawater. Coastal settlements, which include the greater part of Turkmenistan's oil, gas, and chemical enterprises, would also be threatened. Similar environmental problems would certainly affect other Caspian coastal countries as well (e.g., Ragozin, 1995).

The Caspian Sea is unique in yet another respect: It contains about 90% of the sturgeon that produce the lucrative prized black caviar for export to foreign markets. Sturgeon roe is often referred to as "black gold." Today, however, Caspian sturgeon is at risk of extinction from overexploitation by illegal poachers and by destitute fishermen desperately seeking funds to buy food for their families. The sea-level rise, with its destruction of sturgeon spawning grounds, adds yet another threat to the endangered Caspian sturgeon.

Poachers hunt sturgeon only for its caviar. Today, they catch sturgeon directly in the open sea. However, in the early 1960s, prohibition was introduced by the former USSR against catching sturgeon in the open sea. Since that time, catching sturgeon has been carried out in the river deltas. Sturgeon reproduce very slowly: The fish do not spawn for the first time until they reach the age of 20 to 25 years. In 1990, the permissible catch of sturgeon in the USSR was set at 13,500 metric tons. In 1996, permissible (legal) catch was only 1200 metric tons (Rosenberg, 1996).

5 SEA-LEVEL CHANGE AS A GLOBAL PROBLEM

Given the growing concern about, and possible evidence of, global warming, there has been considerable speculation about the potential impacts on coastal areas of a sea-level rise related to global warming. Scientists who participated in the 1995 Intergovernmental Panel on Climate Change (IPCC) Report (IPCC, 1996) suggested that global sea level may well increase by an additional 15 to 70 cm (6 to 27 inches) by the end of the 21st century. The exact amount of rise would depend on the actual increase in global temperatures. Clearly, any additional increase in sea level could have devastating consequences for coastal communities.

All states that border bodies of water, whether along the global oceans or inland seas, should pay attention to fluctuations in sea level as well as to the rise in sea level linked to global warming. Inland seas, for example, can be viewed as living bodies in the sense that they can expand and can shrink. These changes can occur on different time scales: from daily to seasonally, from a year to a decade, or a century, or a millennium. In fact, they fluctuate and change on all these scales. The same can be said of the open oceans, but they tend to fluctuate on much longer time scales than do the inland seas, over periods of many decades and centuries. Such time scales are difficult to factor into the thinking of economic development planners, whose time frames are on the order of years to a few decades at most.

In essence, one can consider the Caspian as a laboratory of sea-level change and its potential societal and environmental consequences. For the Caspian to serve as a true "laboratory," its environmental-monitoring network, which collapsed with the breakup of the Soviet Union, must be restored and maintained by regional cooperation among the Caspian states. Impacts on ecosystems that are managed (farms and pastures) and unmanaged (wetlands, forests, deserts) can be identified. Effective human responses to changes in the coastal zone (both land and sea) can also be identified and assessed; environmental engineering proposals to deal with sea-level changes (such as seawall construction, higher oil platforms in the sea, diversion of water from the Caspian to the drying Aral Sea) can be evaluated for effectiveness, taking into consideration the scientific uncertainties surrounding sea-level fluctuations.

Whether the global climate gets warmer, cooler, or stays as it has been for the last several decades, the level of inland seas will likely continue to fluctuate (the mean ocean level has already gone up by 5 to 6 inches in the twentieth century alone). Societies must learn to cope with both short- and long-term fluctuations. In the Middle Ages, people in the Caspian region were not allowed to settle too close to the sea's shore, under the threat of death. Apparently, leaders were then aware of the dangers that the Caspian's fluctuating levels posed to their citizens. Today's leaders would be well advised to pay attention to traditional wisdom.

REFERENCES

Golitsyn, G. S., Once more about the changes of the Caspian Sea level, *Vestnik AN SSR*, *9*, 59–63, 1989.

Golitsyn G. S., and G. A. McBean, Changes of the atmosphere and climate, *Proc. Russian Acad. Sci. Geogr. Ser.*, *2*, 33–43, 1992.

IPCC, *Climate Change 1995: The Science of Climate Change*, Cambridge University Press, Cambridge, UK, 1996.

Kosarev, A. N., and E. A. Yablonskaya, *The Caspian Sea*, SPB Academic Pub., The Hague, 1994.

Ragozin, A. L., Synergistic effects and consequences of the Caspian Sea level rise, in *Proceedings of the International Scientific Conference: Caspian Region: Economics, Ecology, Mineral Resources*, Geocenter- Moscow, 1995, pp. 120–121 (in Russian).

Rosenberg, I., Catching sturgeon, *Itogi Magazine*, 4 June 1996 (in Russian), 48–50.

CHAPTER 49

ACID RAIN AND SOCIETY

PAULETTE MIDDLETON

1 INTRODUCTION

Acid rain is one of many manifestations of how actions of society can have adverse effects on human health and welfare. Now more than ever before, the breadth of socioeconomic as well as environmental impacts associated with air pollutants, connections among pollutant contributions to these many impacts, and the implications of these connections are being recognized for policy making and development of management strategies.

It can no longer be argued that it is very costly to mediate acid rain and related air quality concerns. Assessments are beginning to suggest that multiple benefits associated with addressing acid rain in combination with other issues outweigh the costs of control of key responsible pollutants. In addition, when innovative strategies, which include market trading and incentives for conservation and use of clean fuels, are initiated, the costs of pollutant management become even lower. As factors that are not easily quantified monetarily are considered more directly in assessments, the benefits become even greater.

Recent analyses show that the implementation of the acid-rain-related part of the 1990 Clean Air Act Amendments has resulted in reductions in acidity in the northeastern United States. Improvements in acid-related impacts have also been suggested. However, projections of future conditions over the next 20 to 50 years suggest that, unless more dramatic steps are taken, the overall burden of harmful pollutants could continue to rise in general in different parts of the country. Dramatic reductions in sulfur oxides, the current main contributors to acidity, alone are probably not enough. Planned reductions in nitrogen oxides may or may not be adequate. Similarly, continued close monitoring, if not increased management, of volatile

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

organic compounds and fine particulates, and reassessment of their importance to acid rain and related concerns will be important over the next few years.

Of course, all of these issues are part of the bigger international picture of energy, environment, and economy. While the United States may be more aggressively and wisely addressing acid rain and related issues here at home, the projections for future fossil fuel use worldwide must be considered for the sake of regional air quality as well as the global climate condition. Capital investment in cleaner technologies such as renewable energy and the promotion of conservation strategies worldwide could bring long-term environmental and economic benefits that far surpass the initial costs. The alternative, continued growth in fossil fuel usage in developing countries, could exacerbate air quality problems that already exist in many of these areas and, in the long term, could cause the same acid rain damages experienced in many parts of North America and western Europe. In addition, it would contribute to adverse long-term carbon-dioxide-induced climate change.

As a contribution to our understanding of the atmospheric pollution problem and our role in the solution, this chapter summarizes:

- Acid rain and its relationship with other major issues
- U.S. response to the acid rain issue
- Current assessment of progress on reducing the effects of acid rain
- Projections and speculation on the future of acid rain

2 ACID RAIN: THE PHENOMENON

The relationships between chemical emissions into the atmosphere and the effects of the chemicals on various ecosystems, human health, and materials are highly complex. Many harmful chemicals (i.e., air pollutants) chemically interact to form other pollutants, which are perhaps even more harmful than the originally emitted chemicals. The most prominent examples of these dangerous chemical products are acid rain, ozone, and aerosols. In addition, many air pollutants are thought to interact in a synergistic fashion to cause even more harm as a group rather than individually. An example of the possible synergism is the hypothesized impact of acid rain and ozone on forest ecosystems. Understanding the causes and effects of acid rain and related air quality issues has become an important mission of atmospheric scientists around the globe.

3 DEFINITION OF ACID RAIN

Acid rain is the general term used to describe the removal, by rainfall, of acidic pollutants from the atmosphere. Acids also can be removed by other forms of precipitation, such as snow or fog. Acid pollutants may also fall as dry particles or gases that form acids when later combined with moisture. The term *acid deposi-*

tion is used to include all the possible forms of acid pollutant removal from the atmosphere, but *acid rain* remains the popular term.

The majority of the deposited acids are nitric acid and sulfuric acid. In some of the more rural regions of the world, organic acids also are important. In very remote areas where the level of acid is low, the "natural background" acid is carbonic acid, which is associated with carbon dioxide in the air. The overall acidity of precipitation also depends on the basic (or alkaline) constituents of the precipitation. Major bases include ammonia and geologic materials, such as dust and fly ash.

Acidity is measured in terms of a pH scale, which is a measure of the log of the hydrogen ion concentration in the precipitation. The scale runs from 0 to 14 with 0 being very acidic and 14 being very alkaline. A midscale value of 7 is considered neutral. A change in 1 pH unit indicates 10-fold increase or decrease in acidity. Unpolluted rain water is considered to have a pH of about 5.6. This acidity is assumed to contain only carbonic acid. In the highly polluted eastern states of the United States., the average acidity of water has a pH between 4 and 5. Even in some remote areas of the world, pH values of 5.2 have been found. These acidity levels suggest that there is a long-range transport of nitrogen and sulfur chemicals.

4 SOURCES OF ACIDITY

Atmospheric acids mainly are produced in the air as a result of complex chemical reactions of the acid precursor gases. Direct emissions of acids such as sulfuric acid, hydrogen chloride, and hydrogen fluoride have been estimated but are not thought to play a significant role in the acidic deposition processes (NAPAP, 1991). Sources of the harmful chemical precursors affecting the acidity of deposited materials can be natural or human caused. The three pollutants of most concern in the acidic deposition process are sulfur dioxide (SO_2), nitrogen oxides (NO_x) and volatile organic compounds (VOCs). Fossil-fuel-based power plants and motor vehicles are the major sources of all of these acid precursor pollutants in industrialized areas of the world (Graedel et al., 1993; Middleton, 1995).

Sulfur gases are primarily emitted from point sources, involving the combustion of coal in particular. Natural sources of sulfur gases include sea spray, volcanoes, and biologic activity. These sources, however, are at least a factor of 10 *less* than the human-caused emission for major industrial areas such as the United States. Nitrogen gases also result primarily from human activities involving fossil-fuel-derived energy use related to transportation and utilities. Major natural sources include soils and lightning and are thought to make up a significant portion of the overall emission totals, more than has been estimated for the sulfur gases. Estimates of these levels, however, are uncertain (NAPAP, 1991). The VOCs that produce the organic acids and influence the chemistry producing sulfuric and nitric acids also come mainly from automobile use. However, for the VOCs, natural production from vegetation can be quite significant. In highly vegetated, low industrialized regions natural sources become the dominant producers of VOCs.

Estimates of alkaline particulate and ammonia emissions indicate that there is a high potential for acid neutralization in some parts of the United States. The estimates, however, are subject to a high degree of uncertainty (NAPAP, 1991). On a global scale, emissions of these important acid neutralizers are among the least well-known chemical emissions (Graedel et al., 1993).

5 EFFECTS OF ACID RAIN

The effects of acid deposition are the subject of continuing controversy. The northeastern United States has experienced the worst reported impacts in the United States (NAPAP, 1991). Severe damages attributed to acid rain also have been documented for parts of western Europe (Graedel and Crutzen, 1989).

The most sensitive systems to acid deposition are poorly buffered lakes and streams. Buffering capacity refers to the availability of alkaline minerals from soil or rocks to neutralize the acids. When minerals are dissolved in a lake, buffering is able to diminish acid effects. However, this buffering ability or alkalinity can be used up with additions of acidic pollutants. Low alkalinity lakes have the greatest potential for damage, since their neutralizing minerals can be quickly depleted.

Vegetation is exposed to wet acidic deposition through rain, snow, and by direct contact with low, acid-laden clouds. There is currently no widespread forest or crop damage in the United States related to these possible pathways. However, cloud acidity, together with a complex combination of other factors (e.g., ozone, soil acidification, climate) contribute to reduced cold tolerance in high-elevation spruce in the eastern United States and in Europe. This can contribute to damage to trees above cloud level during winters with particularly low temperatures.

Adverse effects on forests in other regions of the world are associated with ozone, as is the case with high-elevation pines in California, or they are closely related to localized soil nutrient deficiencies, as is the case with sugar maples in eastern Canada. Acidic deposition may increase leaching rates of important base cations, principally magnesium and calcium, in forest solids and may be a contributing factor in sugar maple decline in some areas.

Generally, controlled experiments on trees and crops have indicated that ozone, at concentrations near ambient levels, adversely affects forests and crops primarily by growth reduction. Other controlled experiments have demonstrated that normal levels of atmospheric sulfur and nitrogen deposition cause no negative direct effects. Some areas actually may benefit through nutrient enrichment by nitrogen and sulfur deposition.

Computer models project that continued acidic deposition could result in long-term deficiencies of nutrients in some soils. However, currently, there is no evidence to indicate that forest health in general is currently affected by nutrient deficiency or will be affected in the next half century.

Air pollution and acidic deposition contribute to the corrosion of metals and deterioration of stone in buildings, statues, and other cultural resources. Although air pollution is an important concern for cultural objects, the magnitude of its effect

on construction materials has been difficult to assess. Many construction materials have protective coatings such as paints; therefore, maintenance and service of protective coatings have an important role in determining the ultimate impact of air pollutants. Paints may also be affected by ambient levels of air pollution.

Another related side effect of acid rain is visibility degradation. Fine particles in the atmosphere containing sulfate, nitrate, and other chemical constituents, which when deposited are associated with acid deposition, cause visibility degradation while in the air. These fine particles have been the major factor in the reduction of visibility in rural and urban areas in the eastern United States since the beginning of the century. In the U.S. West, visibility degradation is being reported in major urban areas and in national parks and wilderness areas.

Direct adverse effects of these pollutants on humans occur largely through the respiratory system. Sensitive populations with existing respiratory or cardiovascular problems, such as those with asthma, are especially susceptible. These effects have been mainly associated with the acid precursor gases and ozone. Studies of the effects of acidic aerosols, composed primarily of nitric acid, ammonium bisulfate, and sulfuric acid, are still relatively new. It has been found that on rare occasions acid levels approach 10 times the long-term mean levels for sulfuric acid. Although substantial uncertainty exists, the body of data raises concern that acidic aerosols alone, or in concert with other pollutants, may be contributing to health effects in exposed populations at current concentration levels.

Human health also can be affected indirectly by pollutants related to acid deposition. People who eat large amounts of fish from acidic lakes or streams may experience exposure to methylmercury in some regions of the country. Drinking water from acidic sources may contain significantly elevated levels of lead. It is unlikely, however, that exposure to humans by this pathway is important, except in isolated cases.

6 SOCIAL RESPONSE TO ACID RAIN

Historically, toxic effects have been observed in populations acutely exposed to high concentrations of air pollutants. As early as the Middle Ages in London prohibitions on coal burning were instituted in response to perceived health effects of mixtures of dust, soot, and fog. The industrial revolution brought the air pollution issue to the United States, where air quality management continued to be considered local in nature into the twentieth century.

The severity of air pollution impacts became very obvious during the London "killer fog" of 1952, when a mixture of particulates, sulfur dioxide, and acidic fog was associated with severe respiratory effects and approximately 4000 deaths. Emergence of air pollution as a public health issue in the 1950s, as a result of this and other deadly episodes, led to the development of federally funded research programs, culminating in the Clean Air Act and in the establishment of the Environmental Protection Agency (EPA) in 1970. These were major stimuli for the establishment of the U.S. National Ambient Air Quality Standards (NAAQS) that today restrict the

atmospheric concentration of pollutants such as sulfur dioxide, nitrogen oxides, and ozone.

Other countries around the world have been developing institutional responses to the threat to human health of air pollution. Air pollution effects on the environment, however, had been slower to be recognized as a serious issue. Acid rain and its ecological effects were first documented in England at the end of the nineteenth century and became regional issues in northwestern Europe and the northeastern United States and eastern Canada only recently—in the late 1960s. During this period and into the 1970s, the mounting anecdotal evidence of the effects of acid rain on aquatic and terrestrial ecosystems launched acid rain as perhaps the first pollution threat to the environment to receive international attention.

The origins of the pressures to regulate acid rain in the United States were primarily twofold. First, Canada protested, lobbied, and publicized its contention that major environmental damage was occurring in its eastern provinces because of acid deposition, and that the major sources of acid precursors were in the United States Second, elected officials and citizens in the eastern and New England states echoed the same concerns, elevating the acid rain controversy to the level of a growing interregional conflict between receptor states and polluting states (Rhodes and Middleton, 1983).

The U.S. responses to these concerns took the form of federal research and eventually control programs. The first step was the Acid Precipitation Act of 1980, which created the National Acid Precipitation Assessment Program (NAPAP). During its first 10 years, the research and periodic assessments conducted by NAPAP improved the understanding of the scientific processes and effects of acid deposition. The monitoring and research conducted in the 1980s and the subsequent integrated assessment completed in 1990 provided the scientific knowledge base for Title IV, the Acid Deposition Control Program, of the 1990 Clean Air Act Amendments.

Title IV is designed to reduce the adverse effects of acid deposition through the reductions in annual emissions of sulfur dioxide (SO_2) and nitrogen oxides (NO_x), the precursors to acid rain. Recognizing that the principal sources of acid rain precursors in the atmosphere are emissions from the combustion of fossil fuels, control measures were initiated to reduce emissions from electric utilities. However, rather than the traditional command-and-control approach to regulation, alternative methods of compliance were allowed. These methods included technological adaptation (e.g., scrubbers, higher-efficiency boilers), fuel-switching, and an innovative SO_2 emissions allowance trading program. This represented the first national effort to use market-based incentives to achieve environmental goals.

Due to the innovative nature of using market-based incentives for environmental regulation, Congress set up a mechanism for checking how well trading was working. As part of this activity, Congress asked NAPAP to assess the costs and economic impacts of the acid deposition control program as well as the effectiveness and benefits associated with the various human health and welfare effects. The effects included visibility, materials, and cultural resources damages and ecosystem effects. NAPAP was also asked to consider the deposition levels needed to protect sensitive

ecosystems. The results of the assessment of Title IV are to be reported to Congress quadrennially, beginning with the 1996 Report to Congress (NAPAP, 1998).

7 CURRENT CONDITIONS

As of the completion of the first report to Congress, several observations have been made regarding the success of Title IV. It appears that the market-based approach has lowered compliance costs. Costs are lower than expected, probably due to a number of factors such as railroad deregulation, technological innovation, and lower operating costs for scrubbers. In addition, all affected utilities have fulfilled the compliance requirements of Title IV. In the first annual reconciliation of allowances and emissions, SO_2 allowances matched or exceeded SO_2 emissions. NO_x reductions have not been as dramatic. This is expected since mandates on NO_x reductions are not in place yet. However, NO_x emissions from all sources in 1995 were 1.5 million tons below 1980 levels. Utilities were responsible for 53% of that reduction.

Statistically significant reductions in acidity and sulfate in precipitation were reported at monitoring sites in the Midwest, mid-Atlantic, and northeastern United States. There is no real evidence of statistically significant decreases in nitrate concentration. Changes in aquatic ecosystems have not yet been detected. However, over the last 15 years, lakes and streams throughout many areas of the United States have experienced decreases in sulfate concentration in response to decreased emissions. While there is some evidence of recovery from acidification in New England, Adirondack lakes continue to acidify, suggesting that additional reductions may be needed in these areas (NAPAP, 1998).

Sulfur and nitrogen deposition has caused adverse impacts on certain sensitive forest ecosystems in the United States, with high-elevation spruce in the eastern United States being most sensitive. Other sensitive forests are apparently not experiencing the same effects in mortality and growth, at least for now, but some of the same processes appear to be slowly occurring.

The leaching of soil nutrients by continued acidic deposition is a gradual process that will eventually impact forest nutrition and growth in many areas. The recent reductions in sulfur should result in some small immediate improvements in sensitive forests, but large improvements will be slow to occur.

Reduced emissions of sulfur oxides are expected to reduce sulfate concentration and its contribution to haze. It is difficult to assess the extent to which recent reductions have contributed to changes in visibility over the past few years since meteorological and other factors determine the overall changes in visibility. Information is needed over the long term.

The recent reductions in SO_x and NO_x emissions are expected to reduce fine particulates and, as a result, lead to improved human health. It is suggested that reduced emissions will lead to a reduction in premature mortality from cardiovascular and respiratory causes and to a dramatic reduction in the number of asthma symptom days.

One difficulty in determining effects at this time is that many impacts have a response times that are longer than the few years since the passage of Title IV. Visibility and acute health effects can be detected on the order of hours to days. Episodic aquatic effects and soil and plant processes in the forest ecosystem respond on the order of days and weeks to months. Chronic human health, chronic aquatic effects, and forest health, on the other hand, indicate response times on the order of years to decades. Effects on forest solid nutrient reserves and effects on materials begin to show up on the order of decades to centuries. These latter effects are more on the order of climate change impacts response times.

The difference in response times, of course, makes an evaluation of actions taken in the early 1990s difficult to quantify. Improvements in health and visibility can serve as indicators of positive change. However, as already noted, even changes in visibility cannot be directly attributed to sulfate reductions alone, since other factors such as meteorological variability play a role in determining visibility changes especially in the humid part of the eastern United States.

8 KEEPING A BROAD BASIS OF ASSESSMENT AND ACTION

To review, acid deposition is an end product of a complex series of interactions among atmospheric chemical species emitted by both natural and human sources. For policy assessment purposes, the most important groups of chemical species are compounds containing sulfur and nitrogen compounds that are emitted from factories, power plants, and automobiles based on fossil-fuel combustion. In addition, volatile organic carbon compounds and fine particles play a role in modulating chemical processes and acidity. Some key compounds remain unchanged in the atmosphere and some are neutralized, but others are oxidized into more acidic forms through a complicated series of chemical, meteorological, physical, and biological interactions.

Decisions about the control of acid deposition must deal with the environmental impacts of estimated future emission levels as well as present levels. Projections depend on many complex and interacting socioeconomic factors. The predictability of how rapidly and to what extent fossil fuels will be replaced by clearer and safer fuels (society may change transportation and other energy use habits) and the interrelationships among countries of the world becoming driving influences for these changes is highly uncertain. Given the demonstrated value of examining multiple causes and effects together, it will continue to be important to keep the base of assessment broad in spite of the uncertainties. The elements of such assessments are illustrated in Figure 1.

On issues related to acid rain, other policy discussions going on throughout North America also illustrate the growing awareness of the interconnections, as well as the need to capitalize on the relationship in developing strategies for the future. For example, EPA, through the Federal Advisory Committee Act (FACA), is leading the development of combined ozone, particulate matter (aerosols), and regional haze implementation program rules and guidance. Other activities at the regional level,

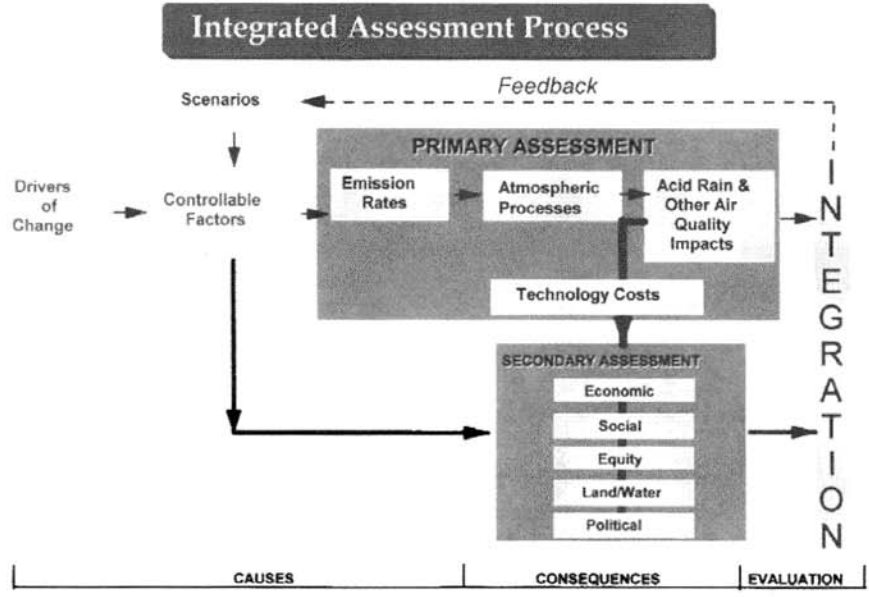


Figure 1 Important elements to consider in an assessment of acid rain in the context of other important environmental, energy, and economic concerns. See ftp site for color image.

such as the Western Governors' Association (WGA) Air Quality Initiative, the Ozone Transport and Analysis Group (OTAG), the Southern Appalachian Mountain Initiative (SAMI), the Southern Oxidant Study (SOS), and the North American Research Strategy for Tropospheric Ozone (NARSTO), are addressing various science and policy issues associated with ozone, particulate matter, and regional haze.

On a continental scale, the Commission on Environmental Cooperation (CEC) is working on North American strategies for addressing transboundary concerns, which include ozone, particulate matter, regional haze, and acid rain along with other hazardous pollutants. Finally, on a much broader scale, the U.S. global climate change program is addressing regional climate assessment for areas throughout the United States. The regional air quality concerns addressed by FACA and the climate concerns to be considered in these discussions are closely linked through the development of aerosols that can influence climate on regional scales as well as produce other problems. On a larger policy implementation level, the two are linked through the development of energy strategies aimed at reducing greenhouse gas emissions and the emission of other more traditionally harmful air pollutants.

All of these programs and approaches share the same fundamental concerns. The role of natural or background processes and the role of chemical interactions in determining the levels of impacts in different regions continue to be fundamental overarching scientific questions. The implementation issues of emissions trading versus pollution prevention versus technological controls are also part of each

aspect of the various debates. Assessments of trade-offs for any decisions made with respect to any of these issues must consider the less quantifiable, and sometimes more uncertain, impacts associated with heath, social impacts and values, equity, and related environmental concerns about water and soil quality as well as air. Given these strong interconnections, it is important to make the best use of research and policy-making resources across organizations addressing acid rain and related issues, where energy and the environment are key factors.

REFERENCES

Graedel, T. E., T. S. Bates, A. F. Bouwman, D. Cunnold, J. Dignon, I. Fung, D.J. Jacob, B.K. Lamb, J.A. Logan, G. Marland, P. Middleton, J.M. Pacyna, M. Placet, and C. Veldt, A compilation of inventories of emissions to the atmosphere, *Global Biogeochem. Cycles*, *7*, 1–26, 1993.

Graedel, T. E., and P. J. Crutzen, The changing atmosphere, *Sci. Am. 261* (Sept.), *58*, 1989.

Middleton, P. Sources of air pollutants, in *Composition, Chemistry, and Climate of the Atmosphere*, Van Nostrand Reinhold, New York, 1995.

Rhodes, S. L., and P. Middleton, The complex challenge of controlling acid rain, *Environment*, *25*, 6–9, 31–38, 1983.

U.S. National Acid Precipitation Assessment Program (NAPAP), *Acidic Deposition: State of Science and Technology, Summary Report of the U.S. National Acid Precipitation Assessment Program*. Washington, DC. 1991.

U.S. National Acid Precipitation Assessment Program (NAPAP), *NAPAP 1996 Report to Congress*, 1998, http://www.nnic.noaa.gov/CENR/NAPAP/NAPAP.96.htm

CHAPTER 50

IMPACTS OF CLIMATE CHANGE

STEWART J. COHEN

1 SCIENCE-POLICY CHALLENGE

The Intergovernmental Panel on Climate Change (IPCC) has concluded that if atmospheric concentrations of greenhouse gases continue to increase, largely as a result of fossil fuel combustion, agricultural practices, and deforestation, average temperatures will increase at a rate much faster than our world has experienced since the last Ice Age. This in itself represents a vision of the future that is substantially different from the past. It challenges long held notions of climate stability, slow rates of climate change (in human terms), and the dominance of natural forces over societal forces in influencing global climate.

Even as new evidence is presented about how civilizations over the last 5000 to 7000 years have been affected by changes in climate (Lamb, 1982), it is recognized that those historic shifts in annual temperatures were only around $\pm$ 1 to 2°C from current global averages. The medieval warm epoch of the tenth to thirteenth centuries was around 0.5 to 1°C above the current global average, while the Little Ice Age of the sixteenth to nineteenth centuries was around 1°C cooler. Although ancient and medieval civilizations may have been less technologically developed than twentieth-century society, modern individuals and nations still have to plan for and adapt to climate. Impacts associated with recent extreme events and El Niño episodes illustrate that despite technological advances, societies in developed and developing countries are still vulnerable to short-term variations in climate (Burton, 1997). There has been a clear upward trend in weather-related costs to insurance companies since the 1960s (Munich Re, 2000), leading to substantial losses by major reinsurance companies such as Lloyd's (IPCC, 1996a). This does not include uninsured losses that may be equal in magnitude to insured losses.

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

A warming of up to 5.8°C during the twenty-first century, a rate of up to 0.5° per decade, would be unprecedented in human history (IPCC, 2001a). Such a change in mean temperatures, with accompanying changes in seasons and probabilities of extreme events, would have direct impacts on land, water, wildlife, and a myriad of indirect impacts on communities, businesses, and governments. If greenhouse gas emissions are not reduced, societies would be faced with the prospect of having to adapt to a new climate that current climate models are not yet able to precisely describe, especially at the regional scale. Adaptation in the twenty-first-century context would be a very different challenge than the one faced by our ancestors, and the costs of adaptation measures are not known.

Identifying societal impacts of a scenario of rapid warming is a complex interdisciplinary research activity that goes beyond assessments of changes in atmospheric processes alone. These changes would be superimposed on changing populations, landscapes, institutions, technologies, and perceptions of resources and environment.

What are the broad dimensions of the societal aspects of global climate change? An outline is presented in Figure 1. This chain of causality also represents a target for a research activity known as *integrated assessment* (IA). Several research groups have attempted to incorporate part or all of this chain into a series of linked models, which collectively have become known as *integrated assessment models* (IAM). The

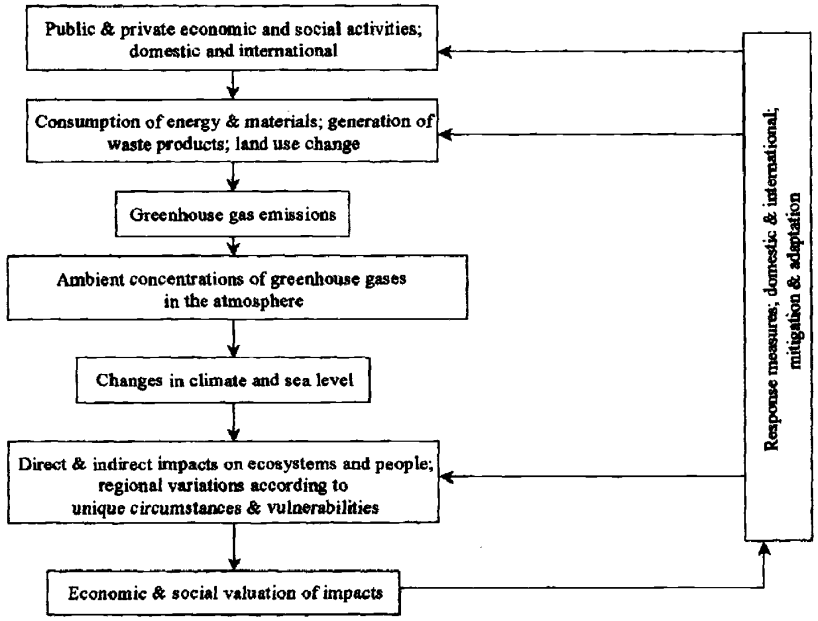


Figure 1 Science-policy dimensions of climate change. Mitigation responses focus on emission reduction or sink enhancement. Adaptation responses focus on reduction of vulnerabilities to climatic events or taking advantage of new climate-related opportunities.

IPCC lists 23 IAMs in its Second Assessment Report (IPCC, 1996b). Other IA techniques include economic models, decision support models, expert judgment exercises, policy exercises, and the use of themes or places as an interdisciplinary platform for collection and analysis of information (Cohen, 1997; Schneider, 1999; Kasemir et al., 1999).

2 NEED FOR INTEGRATED ASSESSMENT OF GLOBAL CLIMATE CHANGE

The outline sketched in Figure 1 suggests that climate change should not be treated in isolation from other environmental concerns, particularly the suite of challenges that are part of global environmental changes resulting from unsustainable development. Sustainability requires living within the carrying capacity of Earth, and this has become a highly political and emotional issue in many places as environment and development come into conflict (e.g., toxic wastes, overfishing, deforestation, desertification). Climate change is connected to many of these concerns, yet climate change has often been treated as a narrowly defined question of atmospheric change and greenhouse gas emissions. Climate change is not only about the meteorology and chemistry of the atmosphere. It is about the underlying human (i.e., economic, political, social) sources of this stress, and the potential victims. Understanding these other components requires a more holistic view of the climate change issue, well beyond consideration of atmospheric science alone.

The advantage of trying an integrated approach is that it represents an explicit attempt to incorporate both physical and human dimensions into research on climate change impacts and responses. For example, a study may indicate a projected change in the potential of a region or country to grow corn and wheat. However, just because there is a change in potential does not mean that farmers, other landowners, communities, businesses, governments, and other stakeholders would agree to a land-use change in response to this change in land capability. Another example is the current debate about measures to reduce greenhouse gas emissions, and the effectiveness of various alternatives such as a "carbon tax," technology transfer from developed to developing countries (through international or binational agreements), or an emissions trading system.

3 METHODOLOGY FOR IMPACT ASSESSMENT OF CLIMATE CHANGE SCENARIOS

Unlike assessments of current and historical events, impact assessment of projected climate change is based on *scenarios*, or plausible pictures of the future. This kind of research can include analysis of past observed events (e.g., the Dust Bowl in the United States during the 1930s), which could serve as societal analogs of a future warm climate (Glantz, 1988), or as climate change analogs superimposed on twenty-first-century society (Rosenberg, 1993), but there are also many studies at regional

and global scales that use climate change scenarios as part of a forward-looking exercise to estimate future impacts.

Case studies of future climate scenarios have often been based on the outputs of climate models (IPCC, 1996a,b; Carter et al., 2000). These models (general circulation models or GCMs) produce estimates of climatic variables on a regular network of grid points for a base case (i.e., current climate), and as an "equilibrium" response to a doubling of carbon dioxide concentrations, or as a "transient" response over 50 years or more of incremental increases in carbon dioxide. Most assessments have taken the difference between the equilibrium or transient simulations and the base case simulation and combined this simulated "change" with baseline climate information from actual observations to produce a scenario (Carter et al., 1994, 2000).

These scenarios of changes in temperature, precipitation, and other elements are used as inputs to other analytical tools that would convert these climate changes to first-order changes in landscapes, ecosystems, renewable resources, and disease rates. Examples include (a) hydrologic models for streamflow and lake levels, (b) crop models for grain yields, (c) fire indicators for forest fire potential, and (d) pest indicators for seasonal ranges of insects. Outputs of first-order impact assessments have been applied to economic models (e.g., timber yields, food production, hydroelectric generation) to estimate impacts in monetary terms (IPCC, 1996a, 2001b).

There have also been attempts to combine these individual estimates into regional or national impact assessments. Scaling up from sectoral to national assessments adds complexity and uncertainty to an already difficult assignment. Economies and societies are composed of many stakeholders whose actions are not easily amenable to modeling. Governments, industries, and individuals may choose various response strategies depending on their knowledge and perceptions of the climate change issue and other forces of concern (e.g., population growth, changes in global trading patterns, and technology).

At the same time, concerns raised by atmospheric scientists about greenhouse gas emissions has led to international negotiations to establish a global strategy to reduce emissions. The main policy instrument, the United Nations Framework Convention on Climate Change (UNFCCC) has been ratified by more than 150 countries, and the emission targets for six greenhouse gases have been tentatively established for more than 30 industrialized countries by a recent agreement known as the Kyoto Protocol (negotiated in December, 1997, but not yet ratified). Critics of this agreement have argued that these targets will cause severe economic losses to most of these countries because, in their view, economic growth is directly linked to growth in energy consumption and hence increased greenhouse gas emissions. Two other concerns are raised as well: (a) global warming may not happen as predicted since there are uncertainties in climate models and their projections may be wrong, and (b) since societies in the past flourished during warm periods (e.g., medieval warm epoch), global warming (if it occurs) would actually be a good thing. Although the IPCC (1995, 1996a,b, 1998, 2001a,b) has published estimates of projected climate change, impacts of climate change scenarios, and impacts of greenhouse gas emission reductions, the rapid policy response has created a substantial new research challenge—determining the impacts of various policy options and compar-

ing these with the costs of doing nothing about emissions. Given the magnitude of the climate change problem and proposed responses to this, the demand for answers will be high.

4 SUMMARY OF CASE STUDIES

Estimates of climate change impacts are summarized for several key sectors in Table 1. In fisheries and health, higher levels of impacts are estimated for developing countries, reflecting their vulnerabilities to climate change. Impacts in agriculture would result from damage due to heat stress, decreased soil moisture, increased incidence of pests and disease, and changes in plant growth cycles; but this could be offset in some circumstances by longer growing seasons and CO_2 fertilization. Some developing countries, however, could experience significant increases in population at risk from hunger. Coastal zone costs are high in many regions, reflecting the growth in built structures in areas vulnerable to sea-level rise and extreme events, as well as land loss itself (e.g., coastal wetlands). Estimates are also available for changes in water supply, wetlands, electricity demand, and some other sectors (IPCC, 1996a,b).

In all cases, scenario estimates are dependent on a variety of assumptions about regional changes in climate (downscaling from GCMs), indirect effects of climate change (e.g., CO_2 fertilization effects), technological change, population growth, changes in infrastructure (e.g., health services in developing countries), and responses of stakeholders to other issues besides climate change (e.g., changing international markets). Table 1 should be considered as a first attempt at determining

TABLE 1 Range of Sectoral Impacts for Different World Regions (2.5°C warming scenario)[a]

Region	Agriculture (% loss in GDP)	Forestry (area lost, km^2)	Coastal Zone (annual protection costs 10^6 U.S.$)	Health (number of deaths, 1000 s)	Fisheries (reduced catch, 1000 ton)
European Union	0.21	52	133	8.8	558
United States	0.16	282	176	6.6	452
FSU	0.24	908	51	7.7	814
China	2.10	121	24	29.4	464
Non-OECD	0.28	334	514	114.8	4,326
OECD	0.17	901	493	22.9	2,503
World	0.23	1,235	1,007	137.7	6,829

GDP, gross domestic product; FSU, former Soviet Union; OECD, Organization of Economic Cooperation and Development (developed countries); non-OECD, developing countries.

Source: Adapted from IPCC (1996b). Table 6.5

TABLE 2 Overall Annual Economic Impacts in Different World Regions for a 2.5°C Warming Scenario (% of current GDP)[a]

Region	IPCC, 1996b	IPCC, 2001
Developed countries	−2.8 to −1.3	−2.8 to 0.3
FSU	−0.7 to 0.3	−0.7 to 11.1
Developing countries	−8.7 to −4.1	−4.9 to 1.8
World	−1.9 to −1.4	−1.5 to 0.1

[a] See Table 1.

Source: Adapted from IPCC, 1996b (Table 6.6) and IPCC, 2001b (Table 19-4).

impacts, and changes in such estimates should be expected as new information becomes available.

Regional impact costs are summarized in Table 2. The range of uncertainty cannot be gauged from the existing literature, nor can the range of estimates provide a confidence interval. Some costs are hidden because of aggregation of communities and nations into large regions and because of lack of information on potential impacts of climate change scenarios on construction, insurance, nontropical extreme events (e.g., midlatitude river floods), transportation, and political institutions.

This work is still in its infancy, and economists have to make assumptions about markets, trading patterns, technological change, adaptation (direct, indirect), and the availability of information. Uncertainties associated with converting impacts into monetary units are due to many factors. One of the most controversial is the cost of health impacts. Is a premature death due to climate change (e.g., heat stress, tropical disease, etc.) worth the same monetary value if it were to occur in a developed or developing country? Since average incomes differ, initial estimates have used a higher "value of a statistical life" for a premature death in a developed country than in a developing country (IPCC, 1996b, Chapter 6). More recent assessments focus on global and regional development trends and their effects on vulnerability to climate change (IPCC, 2001b). These trends may increase adaptive capacity in some circumstances (e.g., community health), but may decrease it in others (e.g., protection of endangered species).

The dilemma of placing a value on life and death is an illustration of the political and social dimensions of the potential impacts of global climate change. Others include (a) the presence of different stakeholders with different visions and goals, (b) issues related to cultural preservation (especially in developing countries and the Arctic), (c) the influence of trade globalization on management of climate-sensitive resources (e.g., fish, water resources), (d) the market value of ecosystems (e.g., wetlands, rain forests, alpine tundra), and (e) intergenerational equity and the choice of discount rates (i.e., a percent change in monetary value due to depreciation, inflation, etc.) used in economic valuation of climate change damages and actions. All of these can influence development choices, including adaptation

choices. Global climate change may have started as a theoretical problem of atmospheric science, but the transmission of information to the public has taken this issue outside of the laboratory and into the real world.

5 LESSONS AND A LOOK AHEAD

Climate impact research is a relatively young endeavor. The term *climate impact assessment* was coined only in the 1970s (Munn, 1979; Kates et al., 1985), and case studies of climate change scenarios have been undertaken for less than 20 years. Advances have been made in incorporating various natural and social science disciplines into the effort, and some important lessons have been learned in the process:

1. There is more than one scenario of future changes for any region, regardless of any scenarios of climate change.
2. There is more than one stakeholder, and one cannot assume that while temperatures change, stakeholders will continue historic patterns of activities. Historic and future decisions result from trade-offs and consensus reached by a broad array of decision makers with different visions.
3. Although there are some preferred options for assessing impacts on *sectors*, there is no single best method available to provide impact assessments for *places*. Parallel sectoral assessments have been the approach of choice in most countries, but some integration exercise needs to become part of this process. Integrated assessment models have become a highly visible option, but these tend to focus more on mitigation and are relatively weak on the impacts and adaptation dimensions, so alternative methods will continue to be important contributors.

Societal aspects of global climate change represent a significant interdisciplinary research challenge, in which atmospheric science and scientists will continue to play an important role. This collaboration will be beneficial for advancing the science as well as for providing better information for stakeholders as they grapple with the human dimensions of this issue.

As consumers of climate information, both from observations and models, researchers on impacts of and responses to climate change uncertainties represent a different type of client than those who work in the atmospheric sciences. This group needs value-added information that can be used as input to other analytical tools, which may or may not have been designed with climate as an explicit input element. Indeed, these tools (e.g., crop yield models, water management models) may be calibrated only to current climate conditions, and their response to climate change scenarios outside their calibration range represents one of many uncertainties in this process. Some of these tools require considerable amounts of data, often at spatial scales too fine to be visible in current GCMs. Impacts researchers are following the progress of downscaling activities with considerable interest (e.g., regional

climate models, statistical techniques), and perhaps this can provide incentive to atmospheric scientists to continue research and development in this area. In the meantime, however, the urgent need for impacts information, as well as for testing and developing methodologies for impact and response assessments, means that currently available methods for scenario construction will continue to be used.

There will also be continued demand for assessments of historic events. Some of these events [e.g., effects of warm years, droughts (El Niño–Southern Oscillation (ENSO)] may serve as possible analogs of future climate change, but important questions arise. How can impacts and costs be attributed? Were these due to the climatic event (at what scale), or to changing vulnerabilities, or both? Did regional or global forces cause the climatic event, and was it consistent with modeled scenarios of future climate changes? The recent increase in insurance losses (Munich Re, 2000) is an example of an observed series of events that would benefit from such an analysis.

Finally, the Kyoto Protocol presents a challenge and an opportunity for new research into the potential impacts of measures to reduce emissions and improve adaptive capabilities. The decision to ratify or not ratify this agreement will be taken on the basis of technical information balanced against preset stakeholder interests. Atmospheric science, in collaboration with researchers from many other disciplines, will play an important role in determining the benefits and costs of various response scenarios.

Past and present impacts have occurred over landscapes and populations that are changing for many reasons. Such changes affect vulnerabilities and costs, as well as responses (e.g., changing land use, insurance programs). Future impacts will be influenced by global economic and institutional changes (e.g., globalization of trade) as well as policy initiatives at various scales. Stakeholders' responses will be determined by attitudes and beliefs about the irnportance of climate change in the context of other challenges. Atmospheric scientists were the first to call attention to global climate change. Now that this has become an international policy concern, there will be greater demands to make scientific views known not only in the traditional refereed literature, but in broader public forums as well.

REFERENCES

Burton, I., Vulnerability and adaptive response in the context of climate and climate change, *Climatic Change*, *36*, 185–196, 1997.

Carter, T. R., M. Hulme, J. E. Crossley, S. Malyshev, M. G. New, M. E. Schlesinger, and H. Tuomenvirta, *Climate Change in the 21st Century—Interim Characterizations based on the New IPCC Emissions Scenarios*, Finnish Environment Institute, Helsinki, 2000.

Carter, T. R., M. L. Parry, H. Harasawa, and S. Nishioka, *IPCC Technical Guidelines for Assessing Climate Change Impacts and Adaptations*, University College, London and Center for Global Environmental Research, Tsukuba, 1994.

Cohen, S. J., Scientist-stakeholder collaboration in integrated assessment of climate change: Lessons from a case study of Northwest Canada, *Environ. Model. Assess.*, *2*(4), 281–293, 1997.

Glantz, M. H. (Ed.), *Societal Responses to Regional Climate Change: Forecasting by Analogy*, Westview Boulder, CO, 1988.

Intergovernmental Panel on Climate Change (IPCC), Contribution of working group I to the second assessment report of the Intergovernmental Panel on Climate Change, in J.J. Houghton, L.G. Meiro Filho, B.A. Callandar, N. Harris, A. Kattenberg and K. Maskell (Eds.), *Climate Change 1995—The Science of Climate Change*, Cambridge University Press, Cambridge, 1995.

Intergovernmental Panel on Climate Change (IPCC), Contribution of working group II to the second assessment report of the Intergovernmental Panel on Climate Change, in R.T. Watson, M. C. Zinyowera, and R. H. Moss (Eds.), *Climate Change 1995—Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analysis*, Cambridge University Press, Cambridge, 1996a.

Intergovernmental Panel on Climate Change (IPCC), Contribution of working group III to the second assessment report of the Intergovernmental Panel on Climate Change, in J. P. Bruce, H. Lee, and E. F. Haites, (Eds.), *Climate Change 1995—Economic and Social Dimensions of Climate Change*, Cambridge University Press, Cambridge. 1996b.

Intergovernmental Panel on Climate Change (IPCC), A special report of working group II, in R. T. Watson, M. C. Zinyowera, and R. H. Moss, (Eds.), *The Regional Impacts of Climate Change: An Assessment of Vulnerability*, Cambridge University Press, Cambridge, 1998.

Intergovernmental Panel on Climate Change (IPCC), *Summary for Policymakers of the IPCC Working Group I Third Assessment Report*, approved in Shanghai, January 2001, available on-line, http://www.usgcrp.gov/ipcc/wglspm.pdf, 2001a.

Intergovernmental Panel on Climate Change (IPCC), Contribution of working group II to the Third assessment report of the Intergovernmental Panel on Climate Change, in J. McCarthy, O. Canziani, N. Leary, D. Dokken, and K. White (Eds.), *Climate Change 2001: Impacts, Adaptation, and Vulnerability*, Cambridge University Press, Cambridge, 2001b.

Kasemir, B., M. B. A. van Asselt, G. Dürrenberger, and C. C. Jaeger, Integrated assessment of sustainable development: Multiple perspectives in interaction, *Int. J. Environ. Pollut.*, *11*(4), 407–425, 1999.

Kates, R. W., J. H. Ausubel, and M. Berberian (Eds.), *Climate Impact Assessment*, SCOPE 27, Wiley, New York, 1985.

Lamb, H. H., *Climate, History and the Modern World*, Methuen, London, 1982.

Munich Re, *Topics—Annual Review of Natural Disasters 1999*, Report 2946-M-e, Munich Reinsurance Group, Munich, Germany, 2000.

Munn, R. E., The framework for a climate impact assessment, in *Carbon Dioxide Issues and Impacts: Proceedings of the Workshop on Energy Carbon Dioxide Issues and Impacts*, Climate Planning Board (F.K. Hare, Chair), Canadian Climate Program (Eds.), 19–22, Atmospheric Environment Service, Downsview, Canada, 1979.

Rosenberg, N. J. (Ed.), Towards an integrated impact assessment of climate change: The MINK study, *Climatic Change*, *24*(1–2), 1–173, 1993.

Schneider, S. H. (Ed.), Topics related to integrated assessment, *Climatic Change*, *41*(3–4), 265–546, 1999.

CHAPTER 51

IMPACTS OF STRATOSPHERIC OZONE DEPLETION

MICHELE M. BETSILL

1 INTRODUCTION

In 1928, Charles Kettering and Thomas Midgley, Jr., scientists with General Motor's Research Corporation in Dayton, Ohio, invented chlorofluorocarbons (CFCs) as a safe alternative to toxic and flammable refrigerants. In 1974, Mario Molina and F. Sherwood Rowland published a paper in *Nature* that linked the use of CFCs to destruction of Earth's stratospheric ozone layer. Today, the international community has made substantial progress toward phasing out the production and consumption of CFCs and in setting up the mechanisms that should help to restore the damaged ozone layer.

Ozone is a molecule consisting of three oxygen atoms (O_3). Approximately 90% of Earth's ozone is found in the stratosphere, the region of the atmosphere that is between 10 and 15 km above Earth's surface. Stratospheric ozone helps regulate Earth's atmospheric temperature structure by absorbing damaging ultraviolet sunlight (UV-B).

Stratospheric ozone depletion is primarily caused by human-made chemicals containing various combinations of chlorine, fluorine, bromine, carbon, and hydrogen. Collectively, these compounds are called halocarbons. They can be divided into compounds containing carbon, chlorine, and fluorine (CFCs) and those containing carbon, bromine, and fluorine (halons). CFCs are used in refrigeration, air-conditioning systems, as foam blowing agents, for cleaning electronic components, and as solvents. Halons are primarily used in fire extinguishers. These compounds break down when they enter the atmosphere; then chlorine and bromine atoms react with ozone to catalyze its destruction (WMO, 1995).

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

The possibility that human-made substances may cause stratospheric ozone depletion was first raised in the late 1960s and early 1970s as part of debates about the development of a large fleet of high-flying supersonic transport (SST) aircraft. These aircraft were to be designed to fly at the speed of sound at an altitude of 45,000 ft (well into the stratosphere). Opponents of the SST program raised concern about the sonic booms the aircraft would cause. Several scientists also suggested that water vapor and nitrous oxide emitted by SSTs would catalyze a chemical process leading to the breakdown of stratospheric ozone molecules. The United States, fueled by public outcry over the sonic boom concern and with full realization of the adverse economic aspects, canceled its SST program before the link between nitrous oxide and ozone depletion could be further investigated (Cagin and Dray, 1993; NAS, 1975). It was Molina and Rowland's 1974 article that prompted the scientific community to explore the role of human-made substances in depletion of the stratospheric ozone layer. In 1995, Rowland and Molina were awarded the Nobel Prize in Chemistry for their work in explaining how the stratospheric ozone layer is destroyed. They shared the award with Paul Crutzen who was recognized for his research linking nitrogen oxides with ozone depletion. This marked the first time the Nobel Prize was awarded for environmental work (Lipkin, 1995).

2 IMPACTS OF STRATOSPHERIC OZONE DEPLETION

Stratospheric ozone depletion was recognized as an environmental problem in need of international attention because it impacts both humans and the natural environment. When stratospheric ozone levels decrease, the amount of UV-B reaching Earth's surface increases (WMO, 1995). The changes in UV-B radiation are highest at high and midlatitudes in both hemispheres while the increases are fairly small in the tropics (UNEP, 1994). Increased levels of UV-B affect human health, the productivity of plant and animal species, as well as the composition of ecosystems.

Impacts on Human Health

Ultraviolet exposure does have some benefits for humans. For example, it initiates the production of vitamin D_3, which is believed to inhibit the growth of tumor cells (UNEP, 1996). However, the balance of evidence indicates that the effects of stratospheric ozone depletion on human health are negative. The major risks include increased incidence of eye diseases, skin cancer, and infectious diseases. When UV-B levels increase, two main organ systems are exposed: the eyes and the skin. The impacts of ozone depletion are mediated through these two systems (Longstreth et al., 1995; UNEP, 1998).

Evidence suggests that increased UV-B radiation exposure may be associated with an increase in the incidence of cataracts, a clouding of the lens of the eye (Longstreth et al., 1995; UNEP, 1998). One review of research on this problem reported that a 1% increase in stratospheric ozone depletion would result in a 0.6 to 0.8% increase in the incidence of cataracts (UNEP, 1994; see also UNEP, 1998).

The most widely known impact of increased UV-B radiation on human health is skin cancer. UV-B radiation damages deoxyribonucleic acid (DNA), which may cause gene mutations and the formation of cancer cells. Some studies estimate that a sustained 10% decrease in average stratospheric ozone concentrations would result in 250,000 new cases of nonmelanoma skin cancer. This is in addition to the 1.2 million cases already reported each year (Longstreth et al., 1995; UNEP, 1996). Many animal species, such as cows, goats, sheep, cats, and dogs, are also at increased risk of developing skin cancer as a result of increased exposure to UV-B radiation (UNEP, 1998).

In an assessment of the effect of the Montreal Protocol and its amendments in protecting the ozone layer, Slaper and his colleagues (1996) concluded these efforts will substantially decrease the growth rate of the incidence of skin cancer over the next century. They found that under a scenario where there were no limits on the production and consumption of ozone-depleting substances, there would be a quadrupling in the incidence of skin cancer by the year 2100. Under the provisions of the Montreal Protocol (a 50% reduction in the production of CFCs by 1999), a doubling in the incidence of skin cancer could be expected in that same period. In contrast, they found the Copenhagen Amendments scenario (a complete phase-out in the production of 21 ozone-depleting substances by January 1, 1996) would result in a 10% increase in skin cancer incidence, peaking in the year 2060. This study lends support to the importance of international efforts to combat stratospheric ozone depletion.

Researchers believe that skin exposure to increased levels of UV-B radiation is also linked to modifications in the human immune system. As a result, the ability of the immune system to respond to certain infectious diseases, such as tuberculosis, leprosy, and Lyme disease, is impaired (UNEP, 1998). Longstreth and her colleagues (1995) predict that higher levels of UV-B will result in increased *severity* and *duration* of diseases such as lupus rather than an increase in their *incidence*.

Impacts on Aquatic Systems

The balance of evidence indicates that increased UV-B radiation can have harmful effects on many species of aquatic organisms and the aquatic systems in which they live (SCOPE, 1993; UNEP, 1998). For example, studies in the Antarctic have linked increased UV-B levels to reduced phytoplankton productivity. Phytoplankton are the basis for the oceanic food chain. UV-B radiation affects the DNA, photosynthesis, enzyme activity, and nitrogen incorporation of phytoplankton. Reduced phytoplankton productivity will likely lead to reduced productivity further up the food chain. It has been estimated that a 16% reduction in stratospheric ozone could lead to a 5% loss of phytoplankton causing a loss of 7 million tons of fish worldwide per year (Häder et al., 1995; UNEP, 1994, 1996). Figure 1 illustrates the effects of UV-B radiation on phytoplankton.

Researchers have also found that enhanced UV-B radiation disrupts the early development of several species of fish, shrimp, and crabs, ultimately affecting their motility (Häder et al., 1995). In damaging aquatic organisms, stratospheric

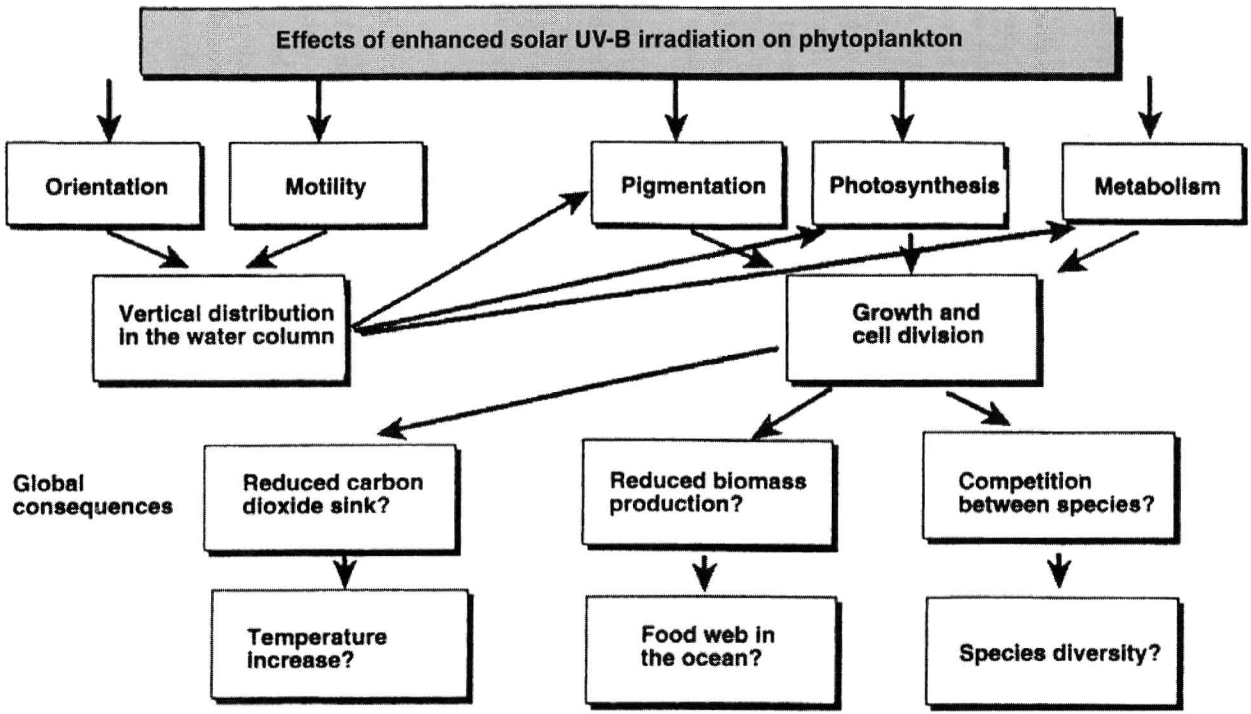


Figure 1 Effects of UV-B radiation on phytoplankton (from Häder et al., 1995, p. 178).

ozone depletion has serious implications for the world food supply. Globally, 30% of the animal protein consumed by humans comes from the oceans. The percentage is much higher in developing countries (UNEP, 1998). These impacts are particularly worrisome in light of the growing world population.

Impacts on Terrestrial Plants and Ecosystems

Scientific understanding of the impact of enhanced UV-B on terrestrial plants and ecosystems is incomplete. The majority of studies have been conducted in growth chambers and greenhouses under controlled conditions, conditions that are often quite different from those experienced in the field. Thus, researchers contend it is necessary to use caution in making generalizations about the impacts of enhanced UV-B on terrestrial plants. The results of existing studies need to be verified under field conditions (Caldwell et al., 1995).

Keeping the limitations of existing research in mind, it is still possible to make some statements about the effect of enhanced UV-B on terrestrial plants. It appears that increased UV-B radiation may have both direct and indirect effects on plants. Some plant species exhibit a reduction in leaf area and/or stem growth when exposed to higher levels of UV-B. In addition, UV-B may also inhibit photosynthesis, damage plant DNA, and alter the time of flowering as well as the number of flowers in some species. The latter has implication for the availability of pollinators and thus the reproductive capacity of plants (Caldwell et al., 1995; UNEP, 1998). The effects of UV-B on plants are not always straightforward but rather depend on the species, the cultivar, and developmental stage of the plants as well as mineral nutrition in the soil, drought, and local air pollutants (Caldwell et al., 1995; UNEP, 1998).

In affecting plants, enhanced UV-B radiation may ultimately lead to changes in entire ecosystems. In nonagricultural ecosystems (e.g., forests and grasslands), the balance of plants may change as some species are less able to respond to increases in UV-B radiation and their productivity declines. At the same time, the productivity of more responsive species will likely increase. The overall species composition of ecosystems will change, as will species interactions and ecosystem dynamics (Caldwell et al., 1995; UNEP, 1998).

Link to Global Climate Change

Increased levels of UV-B radiation may also affect the balance of carbon dioxide (CO_2) into and out of the biosphere, ultimately contributing to the problem of global climate change. For example, phytoplankton absorb CO_2. As discussed above, stratospheric ozone depletion leading to enhanced UV-B is associated with decreased productivity in phytoplankton. This means the reduction of a major sink, resulting in increased levels of atmospheric CO_2 (Häder et al., 1995). In addition, higher levels of UV-B radiation may increase the decomposition rate of nonliving organic matter and reduce photosynthesis, thereby increasing the amount of CO_2 that is emitted into

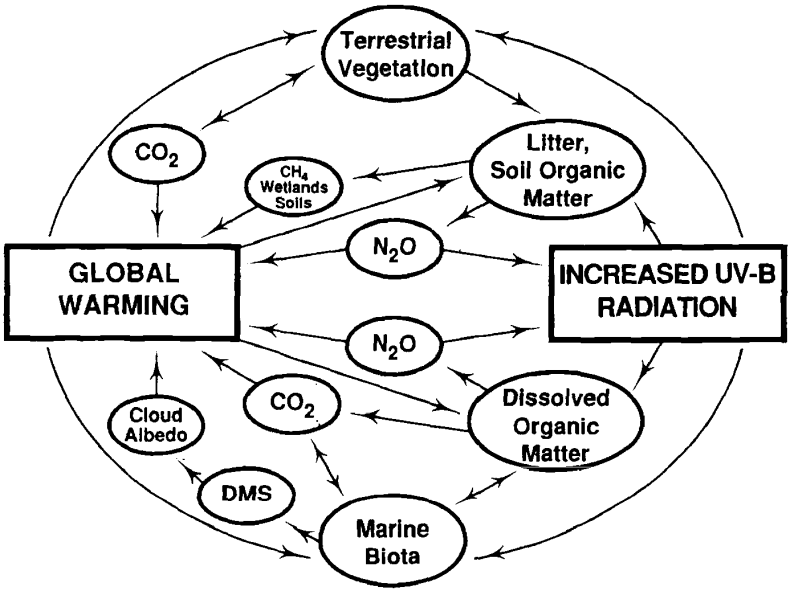


Figure 2 Ways in which increased levels of UV-B are linked to increased levels of CO_2 (from SCOPE, 1993, p. 18).

the atmosphere (SCOPE, 1993; UNEP, 1998). Figure 2 illustrates ways in which increased levels of UV-B are linked to increased levels of CO_2.

3 INTERNATIONAL RESPONSES TO STRATOSPHERIC OZONE DEPLETION

By the early 1980s, it became apparent that combating the problem of ozone depletion would require coordinated international action. Evidence began to mount that ozone depletion was linked to increased rates of skin cancer and that the ozone layer was being depleted at a faster rate than initially imagined (Benedick, 1991; Morrisette, 1989). In 1985, 20 countries signed the Vienna Convention for the Protection of the Ozone Layer, a very general agreement that called upon parties to cooperate in ozone research and to exchange information on the problem of ozone depletion (Roan, 1989; Caldwell, 1990). The Vienna Convention served to establish the recognition of a problem that needed to be dealt with through international cooperation.

The first reports of an "ozone hole" came almost immediately following the signing of the Vienna Convention. Joseph Farman and his colleagues at the British Antarctic Survey station in Halley Bay, Antarctica, who had been monitoring stratospheric ozone over the southern pole since 1957, reported that in September and

October (Austral spring) nearly 60% of the ozone layer over Antarctica was depleted (Farman et al., 1985). Researchers have also found that other regions of the globe experience seasonal decreases in the stratospheric ozone layer. In the midlatitudes, decreases are most severe during the winter/spring months (WMO, 1995).

The 1992 and 1993 Antarctic ozone "holes" were the most severe on record; in some areas, stratospheric ozone was depleted by more than 99%. Researchers believe, however, that these depletions were due at least in part to the eruption of the Mount Pinatubo volcano in 1991. The eruption emitted a substantial amount of sulfate aerosols into the stratosphere, which are believed to enhance the effectiveness of chlorine and bromine as catalysts for ozone destruction (WMO, 1995).

The Montreal Protocol on Substances that Deplete the Ozone Layer was signed by 28 countries in September 1987. The Montreal Protocol established specific measures for countries to control worldwide emissions of several ozone-depleting substances. Specifically, the protocol called for a freeze on the consumption and production of ozone-depleting substances at 1986 levels by 1990, a 20% reduction by 1994, followed by a further 30% reduction by 1999. The agreement entered into force in January 1989. While reports of the ozone hole may have prompted countries to move quickly on the issue of ozone depletion,* the fact that the major CFC producers promised alternatives could be developed in a relatively short time also facilitated international cooperation.

The Montreal Protocol has undergone four major revisions since 1987. The London Amendments to the Montreal Protocol were passed by the Conference of the Parties (COP) at their second meeting in June 1990. These amendments required industrialized countries to completely phase out CFCs by the year 2000 and expanded the number of ozone-depleting substances controlled by the agreement (Benedick, 1991; Parson and Greene, 1995). The protocol was further amended in 1992 at a COP meeting in Copenhagen. The most significant outcome of this meeting was that the phase-out date for CFCs was moved to January 1, 1996 (Parson and Greene, 1995). In September 1997, parties adopted the Montreal Amendments to the Protocol, which strengthened regulation of methyl bromide. The 1999 Beijing Amendment introduced a 2002 phase-out of bromochloromethane and placed controls on the production of and trade in hydrochlorofluorocarbons (HCFCs – a CFC substitute). As of April 2002, there are 183 parties to the Montreal Protocol.

According to a 1995 assessment by the World Meteorological Organization (WMO), the ozone layer is expected to recover by the middle of the next century thanks to these international efforts. CFCs have an atmospheric lifetime of several decades, thus it will take some time before concentrations of these ozone-depleting substances decrease. Researchers expect global UV levels to peak around the turn of the century after which they expect atmospheric concentrations of chlorine and bromine to begin to decline, thereby slowing ozone depletion (UNEP, 1998; WMO, 1995). Figure 3 illustrates the impact of the Montreal Protocol and its amendments on stratospheric ozone levels.

* There is ongoing debate on the role of the ozone hole in the international policy process. For more information, see Benedick (1991), Lambright (1995), Morrisette (1992), Ungar (1995).

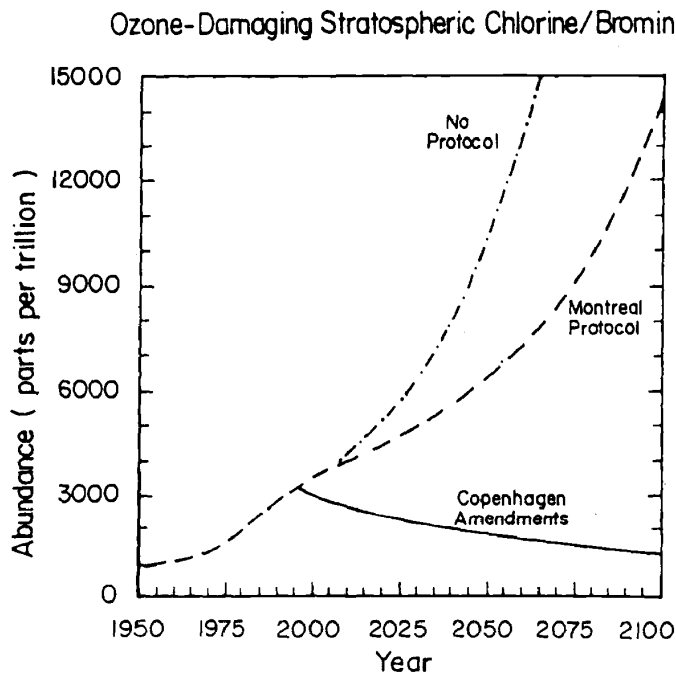


Figure 3 Impact of Montreal Protocol on stratospheric ozone levels (from WMO, 1994, p. 28).

4 REMAINING CHALLENGES IN ADDRESSING STRATOSPHERIC OZONE DEPLETION

Despite the tremendous progress made in addressing the problem of stratospheric ozone depletion, challenges remain. For example, the major replacements for CFCs, HCFCs and hydrofluorocarbons (HFCs), are not totally benign in their effects on stratospheric ozone and may pose other environmental risks. HCFCs do have the potential to deplete stratospheric ozone. However, their atmospheric lifetime is much shorter than for CFCs; thus their ozone-depleting potential is lower than for CFCs and halons (WMO, 1995). In addition, some HCFCs and HFCs react with agents in the atmosphere to produce trifluoroacetic acid (TFA). Once produced, TFA returns to Earth's surface via precipitation. TFA is mildly toxic to both marine and freshwater phytoplankton. There is concern that TFA levels in some areas, particularly those with restricted aquatic outflow, will become toxic (Häder et al., 1995; Schwarzbach, 1995).

Perhaps the greatest remaining challenge in addressing stratospheric ozone depletion is the need to regulate methyl bromide. Bromine, which is derived from methyl bromide, is estimated to be 50 times more efficient at destroying stratospheric ozone

than chlorine (O'Meara, 1996; WMO, 1995). Methyl bromide, which has an average atmospheric lifetime of 1.3 years, accounts for between 5 and 10% of observed ozone depletion. This could increase to 17% if emissions continue to grow at current rates (O'Meara, 1996).

The majority of methyl bromide entering the atmosphere originates from the oceans. This source accounts for 60 to 160 ktons of methyl bromide a year. The primary anthropogenic sources derive from soil fumigation (20 to 60 ktons per year), biomass burning (10 to 50 ktons per year), and exhaust from cars using leaded gasoline (0.5 to 1.5 ktons per year) (WMO, 1995). In addition to its ozone-depleting potential, methyl bromide is also highly toxic to humans and animals. Exposure may result in a variety of symptoms including dizziness, headaches, nausea, respiratory irritation, persistent numbness in the extremities, convulsions, and, in extreme cases, coma or death (Longstreth et al., 1995).

In 1992, parties to the Montreal Convention agreed that developed countries should freeze production of methyl bromide at 1991 levels by 1995. They strengthened the regulation in 1995 and again in 1997, agreeing to a phase out by January 1, 20005.* Developing countries will be required to freeze production of methyl bromide in 2002 based on an average for the years 1995 to 1998. The Montreal Amendments, passed in 1997, established a schedule for phasing out methyl bromide production in developing countries by 2015. Unfortunately, these regulations pertain only to the production of methyl bromide; they do nothing to eliminate the use of methyl bromide.

5 CONCLUSION

The case of stratospheric ozone depletion and the international response to it offer several lessons for dealing with future similar environmental challenges. First, it illustrates that technology may be both a cause and a solution to global environmental problems. The development of CFCs and the use of methyl bromide as a pesticide beginning in the 1930s increased concentrations of ozone-depleting substances in the atmosphere and thus increased the rate of ozone depletion. At the same time, the development of alternative technologies such as HCFCs and HFCs is often seen as the key to motivating countries to take aggressive measures to address the problem of ozone depletion. Without these alternatives, the cost of phasing out the use of CFCs would have been prohibitive.

The ozone case also reminds us that there are no easy trade-offs. While CFCs have high ozone-depleting potential, their alternatives are not entirely environmentally sound. HCFCs and HFCs do have some ozone-depleting potential (though much lower than for CFCs) and may have other negative effects on the environment. Addressing environmental issues often requires making painful choices.

* The United States will phase out production of methyl bromide in 2001 as part of the Clean Air Act Amendments of 1990.

REFERENCES

Benedick, R., *Ozone Diplomacy*, Harvard Press, Cambridge, MA, 1991.

Cagin, S., and P. Dray, *Between Earth and Sky: How CFCs Changed Our World and Endangered the Ozone Layer*, Pantheon Books, New York, 1993.

Caldwell, L. K., *International Environmental Policy: Emergence and Dimensions*, Duke University Press, Durham, NC, 1990.

Caldwell, M. M., A. H. Teramura, M. Tevini, J. F. Bornman, L. O. Bjorn, and G. Kulandaivelu, Effects of increased solar ultraviolet radiation on terrestrial plants. *Ambio*, *24*, 166–173, 1995.

Farman J. C., B. G. Gardiner, and J. D. Shanklin, Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction, *Nature*, *315*, 207–210, 1985.

Häder, D. P., R. C. Worrest, H. D. Kumar, and R. C. Smith, Effects of increased solar ultraviolet radiation on aquatic ecosystems, *Ambio*, *24*, 174–180, 1995.

Lambright, W. H., NASA, ozone, and policy-relevant science, *Res. Policy*, *24*, 747–760, 1995.

Lipkin, R., Ozone depletion research wins Nobel, *Sci. News*, *148*, 262, 1995.

Longstreth, J. D., F. R. de Gruijl, M. L. Kripke, Y. Takizawa, and J. C. van der Leun, Effects of increased solar ultraviolet radiation on human health, *Ambio*, *24*, 153–165, 1995.

Molina, M. J., and F. S. Rowland, Stratospheric sink for chlorofluorocarbons: Chlorine atomic-atalysed destruction of ozone, *Nature*, *249*, 810, 1974.

Morrisette, P. M., The evolution of policy responses to stratospheric ozone depletion, *Nat. Resour. J.*, *29*, 793–820, 1989.

Morrisette, P. M., The Montreal protocol: Lessons for formulating policies for global warming, *Policy Stud. J.*, *19*, 152–161, 1992.

National Academy of Sciences (NAS), *Environmental Impact of Stratospheric Flight: Biological and Climatic Effects of Aircraft Emissions in the Stratosphere*, NAS, Washington, DC, 1975.

O'Meara, M., The next hurdle in ozone repair, *World Watch*, November/December 1996, P. 8.

Parson, E. A., and O. Greene, The complex chemistry of the international ozone agreements, *Environment*, March 1995; pp. 17–20, 35–43.

Roan, S., *Ozone Crisis: The 15-Year Evolution of a Sudden Global Emergency*, J Wiley, New York, 1989.

Schwarzbach, S. E., CFC alternatives under a cloud, *Nature*, *376*, 297–298, 1995.

Scientific Committee on Problems of the Environment (SCOPE), *Effects of Increased Ultraviolet Radiation on Global Ecosystems*, SCOPE Secretariat, Paris, 1993.

Slaper, H., G. J. M. Velders, J. S. Daniel, F. R. de Gruijl, and J. C. van der Leun, Estimates of ozone depletion and skin cancer incidence to examine the Vienna Convention achievements, *Nature*, *384*, 256–258, 1996.

United Nations Environment Programme, (UNEP), *Environmental Effects of Ozone Depletion: Executive Summary*, UNEP, Nairobi, Kenya, 1994.

United Nations Environment Programme, (UNEP), *Environmental Effects of Ozone Depletion: Executive Summary*, UNEP, Nairobi, Kenya, 1996.

United Nations Environment Programme (UNEP), *Environmental Effects of Ozone Depletion: Executive Summary*, UNEP, Nairobi, Kenya, 1998.

Ungar, S., Social scares and global warming: Beyond the Rio Convention, *Soc Nat. Resour.*, *8*, 443–456, 1995.

World Meteorological Organization (WMO), *Scientific Assessment of Ozone Depletion: 1994, Executive Summary*, Global Ozone Research and Monitoring Project-Report No. 37, WMO, Geneva, 1995.

CHAPTER 52

TROPICAL DEFORESTATION AND CLIMATE

ROGER A. SEDJO

1 INTRODUCTION

Tropical forests cover a large portion of the globe's land surface running along the equator, roughly between the Tropic of Cancer and the Tropic of Capricorn. The largest expanse of tropical forest is found in the South American equatorial region, predominantly in the Amazon Basin, but extending up into Central America and down into northern Argentina. Large tropical forests are also found in the equatorial regions of Africa and West Africa and in Southeast Asia, running from India to Malaysia, north into China, and continuing to the islands of the East Indian Archipelago and extending into northeast Australia.

Tropical forests take many forms, largely controlled by variation in rainfall, temperature, and season, but also are affected by soil conditions. The climate of the region between the Tropics of Cancer and Capricorn is uniform in that there is a steady year-round temperature. However, annual rainfall may vary from less than 10 mm along the Peruvian coast to more than 10 m along the Colombian coast only a few hundred kilometers to the north (Terborgh, 1992).

Although tropical forests are often rainforest or wet tropical forest, large areas of dry tropical forests exist in almost all of the regions discussed, covering large areas in South and Central America as well as Africa and, to a lesser extent, Southeast Asia. Tropical forests range from open savannas where rainfall is limited to dense tropical rainforests where rainfall is most abundant. Obviously, the type of tropical forest that occurs in an area depends critically upon the availability of precipitation and moisture. The annual cycle of seasonal change is also an important feature of

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

tropical climates, but the seasons are characterized by variation in rainfall rather than temperature. Evergreen forests occur where there is little or no dry season. Where dry and wet seasons are of approximately equal duration, deciduous forests are the norm.

A unique feature of tropical forests is the broad representation of biodiversity in a small area. Tropical forests contain much, if not most, of the world's biodiversity in the trees and plants that comprise the vegetative system and in the animals, especially anthropoids, that exist in the forest soils, floor, and canopy. Tropical forests, especially wet tropical forests, typically contain far more species of trees, plants, birds, butterflies, and so forth than their temperatc counterparts.

In general, the net primary productivity (NPP), a measure of plant growth given by the net amount of carbon fixed by plants in a given time period, increases the closer the forest is to the equator, although it is moderated by rainfall patterns.

As Table 1 shows, the NPP of tropical forests is substantially greater than that of other types of forest ecosystems, thus indicating higher levels of overall biological growth. This is the case even though tropical soils are often poor in nutrients and minerals as a result of exposure to torrential rainfalls over millennia, which have washed away most of the soluble materials. The high productivity found in these sterile soils is due to the ability of these forests to store the materials in the forest itself—either in living plants or in the litter of dead plants (Terborgh, 1992).

Without human intervention, tropical forest ecosystems are a mixture of living, growing, and dying systems that are periodically impacted by natural disturbances. Natural disturbances, as, for example, when an old tree falls, contribute to this biodiversity by creating numerous unique niches that create a host of varying environments thereby promoting the wide range of biodiversity. The disturbances vary from local disturbances that may open a small hole in the forest canopy that allows sunlight to reach the forest floor thereby stimulating small seedlings, to much broader disturbances. Examples of broader disturbances include those caused by land slides and floods, as well as those caused by forest fires, which occur not only in dry tropical forests but also in wet forests that experience distinct dry seasons, such as those found in some parts of Borneo and Southeast Asia.

TABLE 1 Predicted Net Primary Productivity (NPP) by Forest Ecosystem Types

Ecosystem Type	NPP ($g/m^2/yr$)
Boreal forest	224.20
Temperate conifcr	459.25
Temperate deciduous	361.17
Temperate broadleaf evergreen forest	590.29
Subtropical and tropical	892.74

Source: FAO (1992).

2 HUMAN INFLUENCES IN THE TROPICS

Humans introduce additional complexity to this system by introducing another possible source of disturbance through the process of utilizing forests for their own needs. Few tropical forests are free from the effects of humans. Most tropical forests have been inhabited by humans, many for thousands or tens of thousands of years. Human impacts typically involve the collection of various forest items for food and fiber. Very often humans provide management to increase the future supply of desired outputs. In a subsistence context such management may include modifying the forest to promote certain desired nontimber forest products, e.g., promoting the growth of certain fruit and nut-bearing trees or promoting the growth of young edible bamboo shoots.

In addition, shifting cultivation has been practiced in some tropical forests for millennium. Under this system a small area is cleared and usually burned, in part to provide a nutrient-rich ash. Crops are planted on the site for several years until it loses fertility or the weeds become untenable. The site is left fallow for a period of years, often more than 20, at which time the site is again cleared, burned, and replanted in crops. This cycle appears to be sustainable indefinitely. Thus, even in subsistence and relatively primitive societies, humans have influenced the forest ecosystem. When considering the sustainability of forests, human behavioral influences must be considered as well as natural adjustments.

Forests have also been utilized for millennia by people for building materials and timber. Although there has always been local use and some international trade in tropical timbers, the importance of trade has increased in the post–World War II global economy. Timbers from the West Coast of Africa have been supplemented by much larger volumes entering international trade from the East Indian countries of Malaysia, the Philippines, and Indonesia. This trade has provided these countries with substantial volumes of foreign currencies, which have provided some of the capital that financed their economic development.

3 VALUES OF FORESTS

Tropical forests, indeed all forests, generate a host of values to humans. Human benefits often involve the collection of various forest items for food and fiber such as the various timber and nontimber forest outputs discussed above. In addition, tropical forests provide local and regional ecological services in the form of watershed protection, mitigation of soil erosion, and reduction of downstream flooding. Additionally, tropical forests provide the residence for much of the world's biodiversity, with the majority of species believed to be found in the tropical forest habitat. Finally, tropical forest, together with the rest of the world's forests, provide the majority of the world's biomass that provides a "sink" for carbon, thus mitigating the buildup of carbon in the atmosphere, which is believed to contribute to global warming.

TABLE 2 Recent Estimates of Carbon Flux, Pg C yr^{-1}, from the Tropical Landscape for 1980 and 1990

Source	Year	Range	Average
Detwiler and Hall (1988)	1980	0.4–1.6	1.0
Hall and Uhlig (1991)	1980	0.52–0.64	0.58
Houghton et al. (1987)	1980	0.9–2.5	1.7
Houghton (1992)	1990	1.2–2.2	1.7

Source: Brown et al. (1993, p. 79).

Forests are well recognized as having the potential to affect the forestlands' microclimate. Additionally, forests have an impact on the global climate through their capacity to sequester carbon. Although other terrestrial systems also sequester carbon, forests by far constitute the largest terrestrial carbon pool; tropical forests account for about one-half of all forest area and perhaps a large portion of the total forest biomass (Brown et al., 1993). By holding carbon captive in pools of forest biomass and soils, the amount of carbon in the atmosphere is reduced, thereby mitigating the greenhouse warming provided by the atmosphere.

While it is widely agreed that the preponderance of human-generated releases of carbon into the atmosphere in recent decades has been due to the use of fossil fuels and that the global warming "problem" is largely a fossil fuel problem, it is also recognized that land-use changes play a role. It is generally believed that some of the buildup of carbon dioxide in the atmosphere experienced over the past century or so is due to land-use changes associated with land clearing and deforestation. In recent decades, probably all of the net carbon releases from forests have come from tropical deforestation (since temperate and boreal forests are in approximate carbon balance). An estimate of the carbon releases in recent years is provided in Table 2.

The range of carbon releases from tropical forests is from 0.4 to 2.5 Pg C yr^{-1}. This compares with an estimate of about 6.0 Pg C yr^{-1} total human-generated releases of carbon. Thus, although tropical forest carbon releases are significant, they are well below 50% of total annual releases and probably in the range of 10 to 25% of the total.

4 DEFORESTATION IN THE TROPICS

In 1990 the tropical forests of the world were estimated to cover an area of about 1,756.3 million ha (Table 3), or about 13.4% of the globe's land area, excluding Antarctica and Greenland. This is down from an estimated 1,910.4 million ha in 1981 (Table 3). The annual deforestation rate for this period averaged 0.1540 million km^2 or about 0.8% of the area of tropical forest (FAQ, 1993). The rate of deforestation, however, varied substantially throughout the tropics. Perhaps somewhat surprisingly, the tropical rainforest, the forest type over which the international community

TABLE 3 Estimates of Forest Area and Rate of Deforestation by Geographical Subregions[a]

Continent	Number of Countries	Total Land Area[a]	Forest Area 1980[a]	Forest Area 1990[a]	Annual Deforestation 1981–1990	Rate of Change 1981–1990 (percent per annum)
		million hectares				
Africa	40	2236.1	568.6	527.6	4.1	−0.7
West Sahelian Africa	9	528.0	43.7	40.8	0.3	−0.7
East Sahelian Africa	6	489.7	71.4	65.3	0.6	−0.8
West Africa	8	203.8	61.5	55.6	0.6	−1.0
Central Africa	6	398.3	215.5	204.1	1.1	−0.5
Tropical southern Africa	10	558.1	159.3	145.9	1.3	−0.8
Insular Africa	1	58.2	17.1	15.8	0.1	−0.8
Asia	17	892.1	349.6	310.6	3.9	−1.1
South Asia	6	412.2	69.4	63.9	0.6	−0.8
Continental South East Asia	5	190.2	88.4	75.2	1.3	−1.5
Insular South East Asia	5	244.4	154.7	135.4	1.9	−1.2
Pacific Islands	1	45.3	37.1	36.0	0.1	−0.3
Latin America	32	1650.1	992.2	918.1	7.4	−0.7
Central America	7	239.6	79.2	68.1	1.1	−1.4
Mexico						
Caribbean	19	69.0	48.3	47.1	0.1	−0.3
Tropical South America	7	1341.6	864.6	802.9	6.2	−0.7
Total	90	47,783	1910.4	1756.3	15.4	−0.8

[a] Totals may not tally due to rounding.

Source: FAO (1993).

seems to have the greatest concern, experienced the slowest rate of overall deforestation at 0.6% annually. The highest rates of deforestation were experienced in the upland forests. Both moist and dry upland forests experienced a 1.1% annual rate of deforestation (Table 4). By major region, Central America, including Mexico, experienced the highest rate of deforestation at about 1.4% annually while the Caribbean had the lowest rate at 0.3% annually. Of the large forest formations, continental Southeast Asia had the most rapid rate of deforestation at 1.5% annually while Central Africa had the lowest rate at 0.5% annually.

It is estimated that 90% of tropical deforestation has occurred since 1970 (Skole et al., 1994). If this estimate is correct, the tropical forest of the world at its apex would have covered about 22 million km^2 or about 16.8% of the globe's land surface. Although reduced in size the world's tropical forests still constitute an area equal to that of the whole of South America. Even at the current rate of tropical deforestation, the world's tropical forests would continue to exist through the entire twenty-first century and well into the twenty-second century. Of course, the rate of tropical deforestation will almost surely be changing over time.

TABLE 4 Estimates of Forest Cover Area and Rate of Deforestation by Main Forest[a]

Forest Formations	Land Area (million hectares)	Population Density 1990	Population Growth (1981–1990)	Forest Area 1990		Annually Deforested Area (1981–1990)	
		(inh./km)	(% per year)	(million hectares)	(% of land area)	(million hectares)	(%)
Forest Zone	4186.4	57	2.6	1748.2	42	15.3	0.8
Lowland formations	3485.6	57	2.5	1543.9	44	12.8	0.8
Tropical rainforest	947.2	41	2.5	718.3	76	4.6	0.6
Moist deciduous forests	1289.2	55	2.7	587.3	46	6.1	0.9
Dry deciduous forests	706.2	106	2.4	178.6	25	1.8	0.9
Very dry zone	543.0	24	3.2	59.7	11	0.3	0.5
Upland formations	700.9	56	2.9	204.3	29	2.5	1.1
Moist forests	528.0	52	2.7	178.1	34	2.2	1.1
Dry forests	172.8	70	3.2	26.2	15	0.3	1.1
Nonforest Zone (hot and cold deserts)	591.9	15	3.5	8.1	1	0.1	0.9
Total Tropics	4778.3	52	2.7	1756.3	37	15.4	0.8

[a] Totals may not tally due to rounding.

Source: FAO (1993).

The causes of tropical deforestation are complex and not well understood. The term *deforestation* refers to situations in which the land is more or less permanently converted from forest cover to other cover and/or uses. A common but simplistic view, now largely rejected by analysts familiar with tropical forests, is that tropical deforestation is due to commercial timber logging. Commercial logging in the tropics rarely results in significant direct land conversion; however, as discussed below, it does make indirect contributions to the process of deforestation.

Another common explanation of tropical deforestation is to attribute it to population growth. As populations rise, one would expect that the pressures on the forests might increase. It is also argued that population pressures force a reduction in the fallow period in slash-and-burn regimes increasing pressures to bring more forest land into agricultural use. However, it is difficult to directly link population and economics growth to tropical deforestation, for example, Skole et al. (1994) conclude that population growth alone does not explain tropical deforestation.

Most analysts now believe that most tropical deforestation is driven by human desire for land-use changes, primarily the replacement of forests by agricultural activities. In fact, many governments have currently, or have had in the past, explicit policies to promote the conversion of forests to agricultural uses. In Central and South America, for example, large areas of forest clearing reflect government policy to open forest areas to agricultural settlement. In Brazil, for example, one rationale for promoting migration into the Amazon region, with its attendant deforestation, was to solidity Brazil's sovereignty claims to the region. Other clearing for use as pasture and other agriculture results from spontaneous actions of individuals. In Southeast Asia, forest land was gradually converted to paddy lands in river bottomlands over many years, and more recently conversion has occurred where water development projects, which allow irrigation, have been implemented. Additionally, in Southeast Asia, substantial native forests have been replaced over the years by the introduction of tree crops including rubber, palm oil, coconut, and so forth.

As noted, although commercial logging rarely plays a direct role in deforestation, it often plays an indirect role by providing access to previously inaccessible forest areas. Logging penetrates the dense forest with roads, which then provide access for spontaneous migration. The forest, which is accessible due to the logging roads and is not less dense and impenetrable due to the logging, is now more vulnerable to alternative land-use activities and land-use changes. The improved access makes investments in land clearing, spontaneous and otherwise, more feasible and attractive.

Although governments are often involved in explicitly promoting forest conversion, they are also often involved by the absence of their management of the forests that fall under their jurisdiction. In much of the tropical world today the forests are under the control of the central government, even though in earlier periods they had tended to be under local control. Although the central government has responsibility for the forests, often it is unable or unwilling to exercise effective management control. Thus the forests often become degraded or destroyed by uncontrolled use. In effect, the forests become a type of open access resource over which the responsible authority does not exercise effective control and the users have only illegal or

attenuated use rights that provide little or no incentive for long-term management. More generally, careless use of tropical forest land often signals the absence of clear well-defined property rights and/or effective management, either private or public. With unsecured rights, the incentives are for "cut-and-run" behavior.

5 SIMILARITIES WITH EARLIER DEFORESTATIONS

In many respects tropical deforestation today is not dramatically different from temperate deforestation that occurred one and two centuries earlier. During that period, pressures for land-use change, primarily the demand for new lands for agriculture, resulted in large-scale deforestation of areas of Europe and North America. In the United States much of the forestlands of the eastern seaboard, the south and the Great Lakes states, were converted to croplands and pastures. This same phenomenon had begun earlier in Europe but continued in places well into the early part of the twentieth century. The denuding of the forest landscape was often the result of spontaneous actions but also often reflected govemmental policies. In the United States, for example, the Homestead Act required land clearing as a prerequisite of obtaining land title. For much of North America and Europe the early land clearing has been offset by the renewal of the forest, largely through natural processes. Today, as the work of Kuusela (1994) of the European Forestry Institute has shown, the European forest has reclaimed large areas once deforested. Similarly, in America the forest has reclaimed much of the area deforested in New England (Barrett, 1988), the Great Lakes states, and the south as abandoned agricultural lands regenerated naturally into forest and, more recently, planted forest cover many former tobacco, cotton and other crop lands.

6 TIMBER HARVESTS IN THE TROPICS

Unlike much of the commercial logging in the temperate forest, even today tropical commercial logging almost never involves clear-cutting of the forest. Rather, the usual approach is to select only the trees that are suitable for commercial uses and log those trees, leaving large numbers of live trees in the forest. In past periods, relatively few trees were removed, and those that were felled by hand were commonly transported out of the forest by animals. The forest would be periodically relogged as trees reached desired sizes.

In recent decades logging has involved chainsaws, roads, and equipment. Additionally, larger areas have been logged, reflecting expanded demand. Nevertheless, selective cutting is still almost universally practiced, with only the larger trees of desired species harvested. This practice does not indicate benign motives by loggers but rather reflects the fact that, due to the high diversity of tree species, only a relatively few of the total trees of the tropical forest are suitable for commercial markets. Additionally, unlike many temperate forests that are even-aged (i.e., most all the mature trees are the same age and species), tropical forests typically are

uneven-aged. The diverse mix of tree species and ages makes clear-cutting an unsuitable and uneconomic approach for commercial logging in most of the tropics.

This selective logging approach is also generally conducive to forest regeneration and regrowth. During the period immediately following the logging, the sunlight now reaching the forest floor stimulates the growth of seeds and seedlings, especially of the so-called pioneer species, which include many of the more important timber species such teak, mahogany, and many of the dipterocarp species. Typically the stock of seed and seedlings is adequate; however, this can be supplemented by human activities if required.

The idealized tropical forest management regime regarding logging varies with forest type. In the timber-rich forests of Southeast Asia it is common to follow the logging with a period of 30 to 70 years during which the forest recovers and has time to grow new trees of the desired size. Additionally, existing saplings and medium size trees will continue to grow and, in some cases, accelerate their growth now that the dominant trees are gone. When such an approach is followed, a viable and sustainable forest system can be achieved.

7 RENEWABILITY

Although it is sometimes claimed that tropical forests have difficulty renewing themselves, the evidence is to the contrary. For example, large areas of the American tropics had been in terraces, irrigated agriculture, and agroforestry in the pre-Columbian period but reverted to tropical forests as local populations were decimated by disruptions and disease. For example, Turner and Butzer (1992) argue that "the scale of deforestation, or forest modification, in the American tropics has only recently begun to rival that undertaken prior to the Columbian encounter." Similarly, the great temples of Angkor Watt in Cambodia, Borobodor in Java, and other similar large structures in Southeast Asia, once located in the mist of a high level of human activity, were lost for centuries due to the incursion of tropical forest when human activity declined. Also, the banks of the Panama Canal, which were almost wholly defoliated during the canal's construction in the early 1900, are now covered with lush tropical forests.

8 CONCLUSIONS

A major difference between the prior and current view of tropical deforestation is that the global community is now aware of and concerned for both the preservation of biodiversity and the sequestration of carbon, two global ecological functions that were largely unknown and generated little concern one and one half centuries ago. Another consideration being faced by the global community is increasing awareness that, although deforestation is reversible, species loss is not. Thus, the tropical forests that could be regenerated in the future may not have all of the constituent parts of the original forest. Although there is no guarantee that large amounts of

additional deforestation will be prevented nor that reforestation will follow the pattern of the temperate industrial countries, recent events including the Earth Summit in Rio in 1992 have demonstrated the growing concern over deforestation by the global community and the emerging of a commitment to moderate, if not preclude, its continuance.

REFERENCES

Barrett, J. W., ed; The northeastern region, in *Regional Siliviculture of the United States*, Wiley, New York, pp. 25–65, 1988.

Brown, S., C. Hall, W. Knabe, J. Raich, M. Trexler, and P. Woomer, Tropical forests: Their past, present, and potential future role in the terrestrial carbon budget, in terrestrial biospheric carbon fluxes, in J. Wisniewski and R. N. Sampson (Eds.), Quantification of Sinks and Sources of CO_2, Kluwer Academic, Dordreceht, 1993, pp. 71–94.

Detwiler, R. P., and C. A. S. Hall, Tropical forests and the global carbon cycle, *Science*, *239*, 42–47, 1988.

Food and Agricultural Organization of the United Nations (FAO) *FAO Yearbook: Forest Products, 1981–1992*, FAO Forestry Series No. 27, FAO, Rome, 1992.

Food and Agricultural Organization of the United Nations, (FAO), *Forest Resources Assessment 1990: Tropical Countries*, FAO Forestry Paper 112, FAO, Rome, 1993.

Hall, C. A. S., and J. Uhlig, Refining estimates of carbon released from tropical land-use change, *Canadian Journal of Forest Research*, *21*, 118–131, 1991.

Houghton, R. A., R. D. Boone, J. R. Fruci, J. E. Hobbie, J. M. Melillo, C. A. Palm, B. J. Peterson, G. R. Shaver, G. M. Woodwell, B. Moore, D. L. Skole, and N. Meyers, The flux of carbon from terrestrial ecosystems to the atmosphere in 1980 due to changes in land use: geographic distribution of the global flux, *Tellus*, *39B*, 122–139, 1987.

Houghton, R. A., Tropical forests and climate, paper presented at the International Workshop Ecology, Conservation, and Management of Southeast Asian Rainforests, October 12–14, Kuching, Sarawak, 1992.

Kuusela, K., *Forest Resources in Europe*, Cambridge University Press, Cambridge, 1994.

Skole, D. L., W. H. Chomentowski, W. A. Salas, and A. D. Nobre, Physical and human dimensions of deforestation in Amazonia, *BioScience*, *44*(5), 314–322, 1994.

Terborgh, J., *Diversity and the Tropical Rain Forest*, Scientific American Library, New York 1992.

Turner II, B. L., and K. I. Butzer, The Columbian encounter and land-use change, *Environment*, *34*(8), 16–20, 37–44, 1992.

CHAPTER 53

DESERTIFICATION

R. L. HEATHCOTE

1 INTRODUCTION: ORIGINS OF CONCERNS

Desertification is commonly understood to mean the spread of desertlike conditions. The term came to international attention with the environmental devastation and loss of human and animal life accompanying extensive droughts in West Africa's Sahel region over the period 1968 to 1973, which were captured on televised documentary programs and in the print media. This media coverage likely played an important part in the subsequent extensive United Nations–sponsored international and national aid programs and associated scientific investigations into what appeared to be evidence of long-term environmental deterioration. These activities led to the United Nations Conference on Desertification (UNCOD) in Nairobi, Kenya, in 1977.

In fact, however, concern for evidence of adverse environmental change can be found in the accounts of European explorers in the deserts of Africa, Arabia, and Asia from the late eighteenth century onward. A Danish expedition to Arabia of 1761 to 1767 (Hansen, 1965) and the activities of explorers such as Charles Doughty (1843–1926), Sven Hedin (1865–1952), Baron von Richthofen (1833–1905), and Elsworth Huntington (1876–1947) uncovered evidence of past civilizations amid the desert sands. For example, Huntington's 1907 *The Pulse of Asia* drew upon his own and the earlier explorations to suggest that the evidence of climatic variation and associated environmental deterioration and reduced capacity to support the population in central Asia might have caused the folk migrations that led to the Mongol invasions of southwest Asia and Europe in the thirteenth century.

In the areas of nineteenth-century European colonization, droughts on the Great Plains of North America in the 1880s to 1990s (e.g., Brown, 1948), eastern Australia

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

1895 to 1902 (Heathcote, 1965), and southern Africa in 1918 and the early 1920s (Kokot, 1955) raised concerns for the long-term viability of human occupation. The 1930s saw further concerns for what appeared to be a combination of climatic variability and human mismanagement of the land's resources, with the evidence of the Dust Bowl on the southern Great Plains in 1935 and the first global study of soil erosion (Jacks and Whyte, 1938). In West Africa, Stebbing (1935, 1937) anticipated an advance of the southern edge of the Sahara as a result of overzealous clearance of the forest and savannah vegetation, and Ratcliffe (1938) studied the link between overgrazing and the expansion of the Australian deserts.

Post–World War II reconstruction efforts reflected continuing popular concerns about human use of the land in such writings as Sear's *Deserts on the March* (1949) for the Great Plains, Calder's *Men against the Desert* (1951), and Aubreville's *Climats, Forêts et Desertification de l'Afrique Tropicale* (1949) for African deserts. Pick and Aldis (1944) published concerns for *Australia's Dying Heart*. Lowdermilk (1953) undertook a global review of soil resource mismanagement and repeated earlier calls to reverse the loss of resources.

That such reversals were possible was the claim of the Israeli author, Reifenberg, whose book, *The Struggle between the Desert and the Sown* (1955), documented the efforts of the new nation of Israel to reclaim the deserts as the culmination of a long history of desert advances and retreats in this part of southwest Asia. In 1966 a retired British forester (Baker, 1966) with experience in Kenya and Nigeria and founder of the Men of the Trees Society, published his *Sahara Conquest*, yet another scenario to reclaim the desert, in this case by massive tree-planting schemes.

International scientific interest culminated in the United Nations Arid Zone Research Program, which ran from 1951 to 1971 (UNESCO, 1953). This program brought together scientists from around the world to study the globe's arid areas in an effort to understand both the physical characteristics and the past and present land uses in order to plan better for future management and possible desert reclamation. In effect this program laid the foundations for future United Nations interest in the desertification phenomenon.

This concern over the increasing evidence of environmental deterioration seems to have reflected in part not only the increased scientific evidence of that deterioration, but also a growing moral concern for the global environment and human relations with it. The "environmental movements" of the 1970s and their subsequent "green movements," by focusing upon evidences of resource mismanagement through human ignorance or greed, saw desertification as but one example of the adverse impact of human resource use upon the environment—an impact that had been of concern earlier (Glacken, 1967) and that gained an added ethical dimension (White, 1967; Passmore, 1980; Nash, 1990). Thus, by the 1990s desertification had been consolidated as an international "catch-cry," a political force and a global scientific problem for some of the reasons noted above, but also possibly as a result of the end of the Cold War removing political issues (Driver and Chapman, 1996) that had been competing with environmental ones.

2 DEFINING THE PHENOMENON

Specific definitions of desertification have been many and varied, partly reflecting the fact that it has been seen on the one hand as a *process of environmental deterioration* and on the other as the *product of environmental deterioration*—the devastated environment itself (Glantz and Orlovsky, 1986).

All definitions agree, however, upon certain basic criteria:

- Explicit or implicit evidence of changes in the characteristics of the present environment by comparison with previous characteristics.
- Those changes have resulted in a reduced capacity of the environment to support human life.
- The resultant degraded environment has a desertlike appearance.
- Unless rectified, those changes may continue to further reduce the capacity of the environment to support human life in the future.

To adequately and convincingly identify desertification in those terms, scientists need compatible objective data for a considerable time period for specific areas of the world. Identifying either the processes at work or the end product—the desertified landscape—has proved to be extremely difficult.

The first problem has been the lack of the requisite historical data sets and the assumption of linear trends between those data sets that do exist. The degradation of the soil resource capacity to sustain vegetation or crops (by removal from wind or water erosion, by modification of chemical content through salinization, alkalization, or acidification, or by modification of the soil texture and thus moisture retention capacity by compaction) is recognized to be a significant indicator of desertification. Global descriptions of soil characteristics are still sparse and of varying quality, although formulas for estimates of soil erosion rates are available and have been partly used in the Global Assessment of Human-Induced Soil Degradation (GLASOD). GLASOD was commissioned by the United Nations Environment Programme and was incorporated in the *World Atlas of Desertification* as the most scientifically acceptable measure of desertification (UNEP, 1992).

The other main indicator of desertification is the change in vegetation cover, whether of natural or domesticated plants. Change here may be indicated by reduced vegetation quantity and/or quality (in terms of reduced biodiversity, and/or reduced biomass, and/or productivity), but these indicators are much more difficult to measure. The problem here is the interseasonal and interannual fluctuations in that cover and the extent to which the cover at any given time reflects a linear trend toward increasing or decreasing density or whether it reflects merely cyclical (natural) changes or fluctuations.

In West Africa, where the contemporary concern for desertification originated, there have been serious disagreements among scientists as to the extent of recent vegetation change (Table 1). The differences reflect contrasting assumptions on the

TABLE 1 Conflicting Views of West African Desertification: Forest Loss during the Twentieth Century

	Orthodox View[a] Forest Cover[b]			Revisionist View[a] Forest Cover[b]		
Country	1900	1990s	Loss/Gain (±)	1900	1990s	Loss/Gain (±)
Ghana	9.9	1.6	−8.3	2.5	1.6	−0.9
Cote d'Ivoire	14.5	2.7	−11.8	6.0	2.7	−3.3
Benin	1.1	0.4	−0.7	0.5	0.4	−0.1
Sierra Leone	5.0	0.5	−4.5	0.1	0.5	+0.4
Liberia	6.5	2.0	−4.5	5.5	4.8	−0.7

[a] Sources noted in Fairhead and Leach (1996, p. 189).
[b] Forest cover areas in millions of hectares.
Source: Fairhead and Leach (1996).

historical extent of natural vegetation and the relationships between climate and natural vegetation, differences in the classification of "natural" vegetation, and a failure to recognize historical human-induced deliberate or accidental revegetation of previously sparsely vegetated areas.

The global satellite coverage of the early 1970s onward has improved upon the earlier estimates of vegetation conditions, which had been based upon subjective assessments by individual explorers and other observers. Satellite coverage has provided the base for a Global Vegetation Index—the global vegetation cover averaged over the 1983 to 1990 period as a baseline for subsequent documentation of trends (UNEP, 1992). This of course does not help the debates on changes prior to the 1970s.

A second problem has been the interpretation of the data itself. Faced with what appears to be a desertified landscape in terms of apparently degraded vegetation and/or soils, the scientist must assess whether the landscape is in decline from a more productive past or in process of rehabilitation to a more productive future.

A third problem is the time scale chosen for the analysis of the significance of the environmental changes. While significance on a human scale may be set on scales ranging from interseasonal variations to trends over decades or even a generation (usually seen as 20 to 30 years), significance in ecological terms may range from decades to centuries or millennia, and while these latter scales may be seen as irrelevant for human planning purposes, they may be relevant to any attempt to explain what appears to us to be desertification processes. The basic difficulty remains, however, for the vegetation cover, as for the soil degradation observations, that the implied trend will depend upon the length of time between observations. Oscillations in conditions from whatever cause, which may occur between those observations in time, may not be noticed.

A fourth problem facing definitions of desertification is the fact that there are several interested parties concerned with the phenomenon, some of whom may have an interest in stressing the extent of and dangers from desertification, while others

may be more interested in playing down its significance. For scientists seeking research funding for projects of personal interest, for self-identified victims of the process and their political leaders, for political groups who have identified particular organizations or ideologies as contributing to the process, there may be an incentive to exaggerate the significance of the phenomenon, or at least its natural as opposed to the human causes (Glantz, 1977; Heathcote, 1986). For scientists seeking funding for other research areas, for resource managers accused of exacerbating the desertification processes or seeking to gain time to complete exploitive management strategies, and for political groups anxious to defend specific resource management orthodoxies, there may be an incentive to play down the dangers. There seems to be evidence of both approaches to the phenomenon in the literature (Heathcote, 1980; Garcia, 1981; Watts, 1983; Beinart, 1996; Chapman and Driver, 1996).

A fifth and final problem in defining and specifically in accounting for desertification is that the phenomena are of interest to both natural and social scientists. As a result, the investigations of one group may underestimate the significance of factors of interest to the other. This division has been most evident in the debates about whether desertification stems from natural fluctuations; for example, from climatic variations such as drought periods and the associated vegetation death and soil desiccation and erosion, or from the effects of human activities producing environmental stresses that cannot be sustained in particular locations (Rhodes, 1991; Thomas, 1993).

Bearing in mind such difficulties, the variability of definitions is not surprising. However, the consensus view as adopted by the United Nations has evolved from the definition adopted by the UNCOD in 1977:

> Desertification: The intensification or extension of desert conditions; it is a process leading to reduced biological productivity, with consequent reduction in plant biomass, in the land's carrying capacity for livestock, in crop yields and human wellbeing. (UNCOD, 1977, p. 3)

to that used in the *World Atlas of Desertification*:

> desertification/land degradation is defined as: land degradation in arid, semiarid and dry subhumid areas resulting mainly from adverse human impact. (UNEP, 1992, p. vii)

and further modified as:

> Desertification is land degradation in arid, semi-arid and dry sub-humid areas resulting from various factors including climatic variations and human activities. (UN Convention to Combat Desertification in Countries Experiencing Serious Drought and/or Desertification, Particularly in Africa, June 1994)

Thus, consensus has defined the phenomenon as limited to the drier areas of the world, where highly variable climatic conditions, particularly drought, combined with the adverse impacts of human activities appear to constitute a major threat to the long-term sustainability of support for human occupation in those areas.

3 DOCUMENTING DESERTIFICATION

Since UNCOD in 1977, there have been various estimates of the area desertified. The *World Map of Desertification*, showing areas at risk and prepared for the conference, identified 4.56 billion hectares as "affected or likely to be affected" (UNCOD, 1977, Annex 1). If the extreme deserts were excluded, on the grounds that their condition could not worsen, the area was reduced to 3762 million hectares. In 1984, when the progress of the UN Plan of Action to Combat Desertification was reviewed, the area was reduced to 3.48 billion hectares (Tolba, 1984). Dregne and colleagues estimated 3.56 billion hectares in 1991 (Dregne et al., 1991), while the latest UN estimate is 1.04 billion hectares (UNEP, 1992).

In fact, these varying figures are not strictly compatible, since they refer to drylands defined in different ways at the different times and to the inclusion of areas suffering vegetation degradation but that may not have been suffering soil degradation. At least the most recent UN estimate (UNEP, 1992) does include estimates of soil degradation based upon GLASOD.

The 1992 estimates (Table 2) provide basic information on soil rather than vegetation degradation and while, therefore, marginally more acceptable, must still be viewed with caution. The message is that over 15% of the global soils appear to be degraded, over half of that area (53% of the 1.04 billion hectares) is located within the global drylands. The largest degraded areas are in Asia and Africa, with 38 and 25%, respectively, of the global total, and the proportions of the global degraded drylands are similar—36 and 31%, respectively. Interestingly, the largest proportion of regional drylands desertified is in Europe (33%), but all regions have at least 10% of their drylands showing desertification.

TABLE 2 Global Soil Degradation, c. 1992

Region	Soil Degradation (million hectares)					
	In Susceptible Drylands (%)[a]		In Other Areas (%)[b]		Total Degraded Area (%)[c]	
Africa	319.4	24.8	174.8	10.4	494.2	25.2
Asia	370.3	22.1	376.6	14.6	746.9	38.0
Australasia	87.5	13.2	15.4	7.0	102.9	5.2
Europe	99.4	33.1	119.4	18.3	218.8	11.1
North America	79.5	10.9	78.7	5.4	158.2	8.1
South America	79.1	15.3	164.3	13.1	243.4	12.4
Total global	1035.2	20.0	929.2	11.9	1964.4	100.0

[a] Percentage of total area of regional susceptible drylands.
[b] Percentage of total of regional other areas.
[c] Percentage of global total degraded area.
Total degraded area of 1964.4 million hectares is 15.1% of global land area.
Source: UNEP (1992, p. 25).

TABLE 3 The Most Serious Soil Degradation Areas in Drylands, c. 1992

Region	Soil Degradation Areas (strong and extreme) (million hectares)	Percentage of Global Total Areas (strong and extreme) (%)
Africa	74.0	53.8
Asia	43.7	31.8
Australasia	1.6	1.1
Europe	4.9	3.5
North America	7.1	5.2
South America	6.3	4.6
Total	137.6	100.0

Source: UNEP (1992, p. 13).

The most serious dryland soil degradation and implied desertification is identified in Table 3. Using the worst two categories of soil degradation identified (strong and extreme degradation), Asia and Africa again dominate the scene, although it is Africa's turn to lead with almost 54% of the world's seriously degraded dryland areas compared to Asia's 32%. By comparison, the other regions each contain less than 6% of the global areas.

Maps of the desertified areas of the world, regardless of the definitions used, all show the most intensive areas of desertification to lie on the humid edge of the drylands (UNCOD, 1997; Dregne et al., 1991; UNEP, 1992). This is usually explained as the result of the encroachment of agriculture, with its implied soil disturbance, into areas traditionally given over to purely livestock grazing, together with increasing pressure on vegetation resources for fuel and building materials from increased human populations. This somewhat simplistic view, however, has been criticised as noted below.

4 EXPLAINING DESERTIFICATION

There is an extensive literature devoted to alternative explanations for the desertification phenomenon. Essentially, the arguments fall into three camps. First are those that claim natural processes, specifically climate change or variability—mainly through drought, and feedback linkages between the characteristics of the ground surface and air temperatures, are the cause (Bryson and Murray, 1977); second are those that place the blame squarely on human mismanagement of the environment, mainly through human resource demands that exceed the capacity of the environment to supply in the long term (Sinclair and Fryxell, 1985); and third are those that see a mix of both natural processes and human activities combining and interacting to create an unstable environment (Hare et al., 1977).

In the *World Atlas of Desertification* (UNEP, 1992) the causes of desertification were listed as entirely human derived (Table 4) and no specific listing of natural

causes was provided. Implicit, however, were links between human activities and the natural ecosystems. As identified, *deforestation* and *overgrazing* imply the reduction of vegetative cover resulting in less protection for the soils from solar insolation and increased wind speeds, leading to increased evaporation and potential wind and water erosion, along with possible increases in groundwater levels and soil waterlogging from reduced vegetation soil moisture requirements. *Agricultural* activities include the plowing up of fragile soils, thus baring them for wind erosion; attempts to grow crops in areas with insufficient soil moisture leading to crop failure and soil exposure to further erosion; and the excessive application of irrigation water leading to the buildup of salts in the soil (salinization).

Overexploitation implies excessive use of vegetation for fuel or building materials that reduces the capacity of the vegetation to reproduce or recycle essential nutrients to maintain soil fertility. *Bioindustrial* impacts imply contamination from pollution sources and are usually associated with intensive land uses outside the drylands, hence the relatively small area shown.

There is no doubt that climate variability, particularly in terms of precipitation trends over decades, has had measurable impacts upon the success of human occupation of the drylands of the globe. Some of the most telling evidence comes from the Sahel (Kates, 1981; Hare, 1983; Nicholson, 1994). In such locations seasonal climates resemble desert climates (low precipitation, high evaporation, and high solar insolation) so that any human resource management involving reduction of protective vegetation cover at such time would run the risk of permanently damaging the environment (Aubreville, 1949; Bullock and Le Houerou, 1996).

Such damage might be interpreted as the result of longer-term climate change, such as a trend toward increasing aridity, and labeled desertification. However, dryland ecosystems demonstrate considerable ability to recover from periods of desiccation so that whether or not desertification is identified could depend upon the time period chosen for the study, as suggested earlier. Indeed, the edge of the Sahara Desert defined in terms of vegetation cover has been identified from satellite imagery to have shifted 240 km south between 1980 and 1984, but after several annual retreats and advances was by 1990 only 130 km south of the 1980 location (Tucker et al., 1991). Such variation suggests the importance of short-term fluctuations rather than long-term climatic changes.

Certainly, the clearance of vegetation for cropping or through excessive livestock grazing may change the albedo (reflectivity) of the ground surface and, thus, increase the air temperatures, thereby increasing evaporation and creating droughtlike conditions that encourage soil erosion as a result of increased wind speeds. Significantly, most desert reclamation techniques involve attempts to establish new vegetation cover to directly protect the soil and to act in part as wind breaks (Baker, 1966; Zhu and Liu, 1981).

Table 4, however, omits mention of some of the less obvious but possibly equally important factors, such as the history of colonialism and recent economic development both of which have been identified as relevant to the spread of desertification. Most colonial powers in Africa, for example, introduced new tax systems that forced a monetary economy upon a previously subsistence pastoral or agricultural commu-

TABLE 4 Causes of Desertification

Attributed Cause	Area Affected (million hectares)	Percentage of Total Desertified
Deforestation	578.6	29.4
Overgrazing	678.7	34.5
Agricultural	551.6	28.1
Overexploitation	132.8	6.8
Bioindustrial	22.7	1.2
Total	1964.4 m.ha.	100.0%

Source: UNEP (1992, p. 25).

nity, requiring extra numbers of livestock to be grazed for cash sale to pay taxes, or food crops on the better soils to be replaced by cash crops and necessary food production to be pushed into areas climatically more vulnerable or possessing poorer soils more prone to erosion (Franke and Chasin, 1980; Garcia, 1981; Watts, 1983; Baker, 1984; Macdonald, 1986; Morgan and Solarz, 1994). More recently, improved veterinary services and permanent water supplies developed with foreign aid, which replaced traditionally limited seasonal supplies, allowed larger herds to be carried all year, and this often led to overgrazing of the ranges (Glantz, 1977).

In addition, the increase of human populations in the drylands threatened by desertification, from 57 millions in 1977 to 135 millions by 1984 (UNEP, 1992, p. iv), and particularly the rising populations in Africa (Caldwell and Caldwell, 1990) have brought increased pressure on the environment and, in association with the periodic devastation from civil strife and warfare, have no doubt contributed to the desertification process (Glantz, 1987).

5 FUTURE OF DESERTIFICATION

Despite over 20 years of international efforts, the complexity of the factors involved in desertification has meant that those efforts to reverse the trends have had limited success. Reviews of the results of the 1977 UNCOD initiative were not particularly impressive (Mabbutt, 1987; Odingo, 1992; Rapp, 1987; Spooner, 1987). And despite ongoing research and publications by the United Nations Environment Programme through its *Desertification Control Bulletin*, which documents attempts to halt desertification, the debate about definitions cannot hide the fact that the phenomenon continues to be extensive and locally increasing in the area it affects.

A new UN "Convention to Combat Desertification in those Countries Experiencing Serious Drought and/or Desertification, Particularly in Africa," was agreed to in 1994 and came into force in December 1996. The new emphasis is on support for local schemes to rehabilitate desertified areas or to reduce the potential for their expansion. This shift in scale holds better promise for the future since the specific causes of desertification are more often related to local economic, social, and poli-

tical events in the context of the seasonal weather rather than to broad changes in climatic patterns.

Having said that, however, global warming climate scenarios (in essence forecasts) could result in increased aridity in the drylands, with associated increases in natural soil erosion, even without human interference (Bullock and Le Houerou, 1996). Such a scenario does not offer much hope for reversing desertification in the drylands in the future, unless the pressure imposed by human resource uses can, itself, be reduced.

REFERENCES

Aubreville, A., *Climats, Forêts et Dèsertification de l'Afrique Tropicale*, Societès d'Èditions Geographiques, Maritimes et Coloniales, Paris, 1949.

Baker, R., *Sahara Conquest*, Butterworth, London, 1966.

Baker, R., Protecting the environment against the poor: The historical roots of the soil erosion orthodoxy in the Third World, *Ecologist*, *14*, 53–60, 1984.

Beinert, W., Environmental destruction in Southern Africa, in T. S. Driver, and G. P. Chapman (Eds.), *Time-Scales and Environmental Change*, Routledge, London, 1996, pp. 256–268.

Brown, R. H., *Historical Geography of the United States*, Harcourt Brace, New York, 1948.

Bryson, R. A., and R. J. Murray, *Climates of Hunger*, University of Wisconsin, Madison, 1977.

Bullock, P., and H. Le Houerou, Land degradation and desertification, in R. T. Watson, M. C. Zinyowera, and R. H. Moss (Eds.), *Climate Change 1995, Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analyses*, Cambridge University Press, Cambridge, 1996, pp. 177–189.

Calder, R., *Men Against the Desert*, Allen and Unwin, London, 1951.

Caldwell, J. C., and P. Caldwell, High fertility in sub-Saharan Africa, *Sci. Am.*, *262*(5); 82–88, 1990.

Chapman, G. P., and T. S. Driver, Time, mankind, and the Earth, in T. S. Driver and G. P. Chapman (Eds.), *Time-scales and Environmental Change*, Routledge, London, 1996, pp. 1–24.

Driver, T. S., and G. P. Chapman, (Eds.), *Time-Scales and Environmental Change*, Routledge, London, 1996.

Dregne, H., M. Kassas, and B. Rosanov, A new assessment of the world status of desertification, *Desertif. Contr. Bull.*, *20*, 6–18, 1991.

Fairhead, J., and M. Leach, Reframing forest history, in T. S. Driver, and G. P. Chapman (Eds.), *Time-Scales and Environmental Change*, Routledge, London, 1996, pp. 169–195.

Franke, R. W., and B. H. Chasin, *Seeds of Famine: Ecological Destruction and the Development Dilemma in the West African Sahel*, Landmark, Montclair, NJ, 1980.

Garcia, R. V., *Drought and Man: The 1972 Case History*, Vol. 1: *Nature Pleads Not Guilty*, Pergamon, New York, 1981.

Glantz, M. H. (Ed.), *Desertification: Environmental Degradation in and around Arid Lands*, Westview, Boulder, CO, 1977.

Glantz, M. H. (Ed.), *Drought and Hunger in Africa: Denying Famine a Future*, Cambridge University Press, London, 1987.

Glantz, M. H., and N. S. Orlovsky, Desertification: Anatomy of a complex environmental process, in K. A. Dahlberg, and J. W. Bennett (Eds.), *Natural Resources and People: Conceptual Issues in Interdisciplinary Research*, Westview, Boulder, CO, 1986, pp. 213–229.

Hansen, T., Arabia Felix, in *The Danish Expedition of 1761–1767*, J. McFarlane, and K. McFarlane (Trans.), Readers Union, Collins, London, 1965.

Hare, K., *Climate and Desertification: A Revised Analysis*, World Meteorogical Organization (WMO) World Climate Program Publication 44, WMO Geneva, 1983.

Hare, F. K., R. W. Kates, and A. Warren, The making of deserts: Climate, ecology and society, *Econ. Geogr.*, *53*(4), 332–346, 1977.

Heathcote, R. L. (Ed.), *Perception of Desertification*, United Nations University, Tokyo, 1980.

Heathcote, R. L., *Back of Bourke: A Study of Land Appraisal and Settlement in Semi-Arid Australia*, Melbourne University Press, London, 1965.

Heathcote, R. L., Climate and famine: Differing interpretations of the linkages, *Austral. Overseas Disaster Response Org. Newslett.*, *4*(4), 6–8, 1986.

Huntington, E., *The Pulse of Asia*, Yale University Press, New Haven, CT, 1907.

Jacks, G. V., and R. O. Whyte, *The Rape of the Earth: A World Survey of Soil Erosion*, Faber and Faber, London, 1938.

Kates, R. W., Drought in the Sahel: Competing views on what really happened in 1910–14 and 1968–74, *Mazingira*, *5*(2), 72–80, 1981.

Kokot, D. F., Desert encroachment in South Africa, *Afr. Soils*, *3*(3), 404–409, 1955.

Lowdermilk, W. C., *Conquest Land Through 7,000 Years*, United States Department of Agriculture Soil Conservation Service Agricultural Information Bulletin No. 99, U.S. Government Printing Office, Washington, DC, 1953.

Mabbutt, J. A., A review of progress since the UN conference on desertification, *Desertif. Control Bull.*, *15*, 12–23, 1987.

Macdonald, L. H., *Natural Resources Development in the Sahel: The Role of the United Nations System*, United Nations University, Tokyo, 1986.

Morgan, W. B., and J. A. Solarz, Agricultural crisis in Sub-Saharan Africa: Development constraints and policy problems, *Geogr. J.*, *160*(1), 57–73, 1994.

Nash, R., *The Rights of Nature: A History of Environmental Ethics*, University of Wisconsin Press, Madison, 1990.

Nicholson, S. E., Variability of African rainfall on interannual and decadal time scales, in D. Martinson, K. Bryan, M. Ghil, T. Karl, E. Sarachik, S. Sorooshian, and L. Talley, (Eds)., *Natural Climatic Variability on Decade-to-Century Time Scales*, National Academy of Sciences, Washington, DC, 1994.

Odingo, R. S., Implementation of the Plan of Action to Combat Desertification (PACD) 1978–1991, *Desertif. Control Bull. 21*, 6–14, 1992.

Passmore, J. Man's responsibility for nature, in *Ecological Problems and Western Traditions*, 2nd ed., Duckworth, London, 1980.

Pick, J. H., and V. R. Alldis, *Australia's Dying Heart: Soil Erosion and Station Management in the Inland*, 2nd ed., Melbourne University Press, Melbourne, 1944.

Rapp, A. Reflections on desertification 1977–1987: Problems and prospects, *Desertif. Control Bull.*, *15*, 27–33, 1987.

Ratcliffe, F. N., *Flying Fox and Drifting Sand*, 2nd ed., Angus and Robertson, Sydney, 1938.

Reifenberg, A., *The Struggle between the Desert and the Sown*, Jewish Agency, Jerusalem, 1955.

Rhodes, S. L., Rethinking desertification: What do we know and what have we learned? *World Devel.*, *19*(9), 1137–1143, 1991.

Sears, P. B., *Deserts on the March*, Routledge and Kegan Paul, London, 1949.

Sinclair, A. R. E., and J. M. Fryxell, The Sahel of Africa: Ecology of a disaster, *Can. Zool. J.*, *63*, 987–994, 1985.

Spooner, B., The (apparent) paradoxes of desertification, *Desertif. Control Bull.*, *15*, 40–45, 1987.

Stebbing, E. P., The encroaching Sahara: The threat to the West African colonies, *Geogr. J.*, *85*, 506–524, 1935.

Stebbing E.P., The forests of West Africa and the Sahara: A Study of modern conditions, *Geogr. J.*, *90*, 550–552, 1937.

Thomas, D. S. G., Sandstorm in a teacup? Understanding desertification, *Geogr. J.*, *159*(3), 318–331, 1993.

Tolba, M. K., A harvest of dust? *Environ. Conserv.*, *11*(2), 1–2, 1984.

Tucker, C. J., H. E. Dregne, and W. W. Newcomb, Expansion and contraction of the Saharan Desert from 1980 to 1990, *Science*, *253*, 299–301, 1991.

United Nations Conference on Desertification (UNCOD), *World Map of Desertification at a Scale of 1:25000000*, FAO and UNESCO, New York, 1977.

UNCOD, *Convention to Combat Desertification in Those Countries Experiencing Serious Drought and/or Desertification, Particularly in Africa*, Secretariat of the UN Convention to Combat Desertification, United Nations, New York, 1994.

United Nations Conference on Desertification (UNCOD), *Round-Up, Plan of Action and Resolutions*, United Nations, New York, 1978.

United Nations Environmental Programme (UNEP), *World Atlas of Desertification*, Edward Arnold, London, 1992.

United Nations Economic Scientific and Cultural Organisation (UNESCO), *Arid Zone Research*, Series No. 1 to No. 29, UNESCO, Paris, France, 1968.

Watts, M., *Silent Violence*, University of California Press, Berkeley, CA, 1983.

White, L., The historical roots of our ecologic crisis, *Science*, *155*(3767), 1203–1207, 1967.

Zhu, Z., and S. Liu, Desertification and desertification control in northern China, *Desertif. Control Bull.*, *5*, 13–19, 1981.

CHAPTER 54

IMAGINABLE SURPRISE

STEPHEN H. SCHNEIDER

1 INTRODUCTION

Although unexpected events are often considered "surprises," it is often the case that many events are anticipated by at least some observers. Thus, an unexpected event may be labeled more appropriately as an "imaginable surprise," which is defined as an event or process that departs from the expectations of some definable community. Imaginable surprise is a concept related to, but distinct from, risk and uncertainty. Risk is typically defined as the condition in which the event and the probability that it will occur is known. However, risk almost always is accompanied by a certain degree of uncertainty. Uncertainty remains a difficult concept to define or codify but is usually defined as the condition in which the event, process or outcome is known, but the probability that it will occur is not known. Two basic options are appropriate in the face of uncertainty: (1) reduce the uncertainties through data collection, research, modeling, simulation techniques, etc. and (2) manage or integrate uncertainty directly into the decision-making or policy-making process. When uncertainties are large, a strategic approach that considers a wide range of possible outcomes, including low-probability events, may be a more appropriate way to manage uncertainty. It may be possible to identify "imaginable conditions for surprise" when the conditions that might induce surprises are known even though the actual surprise events are not.

Decision makers at all levels (individuals, firms, and local, national, and international governmental organizations) are concerned about reducing their vulnerability to (or the likelihood of) unexpected events or surprises. After briefly and selectively reviewing the literature on uncertainty and surprise, I introduce a definition of surprise that does not include the strict requirement that it apply to a wholly unexpected outcome, but rather recognizes that many events are often anticipated by

Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impacts, Edited by Thomas D. Potter and Bradley R. Colman.
ISBN 0-471-21489-2

some, even if not most observers. Thus, an imaginable surprise is defined as an event or process that departs from the expectations of some definable community, yet is a concept related to, but distinct from, risk and uncertainty. Therefore, what gets labeled as "surprise" depends on the extent to which what happens departs from community expectations and on the salience of the problem. Impediments to overcoming ignorance range from the need for more "normal science" to phenomenological impediments (e.g., inherent unpredictability in some chaotic systems) to epistemological ignorance (e.g., ideological blocks to reducing ignorance). The substantive focus in this chapter will concentrate on the theme of global change. Examples of imaginable surprises in the global change context are presented, as well as their potential salience for creating unexpectedly high or low carbon dioxide emissions. Improving the anticipation of surprises is an interdisciplinary enterprise that should offer a skeptical welcoming of outlier ideas and methods.

Strictly speaking, a surprise cannot be anticipated; by definition it is an unexpected event. Potential climate change, and more broadly global environmental change, is replete with the truly unexpected because of the enormous complexities of processes and interrelationships involved and our insufficient understanding of them (such as coupling ocean, atmosphere, and terrestrial systems). However, risk, hazard, and related research demonstrate repeatedly that the event, process, or outcome registered as a surprise by the community in question was frequently known or forecast by others or the same event was knowable within the competing frameworks of understanding (Darmstadter and Toman, 1993).

2 UNCERTAINTY

Much of the current work on surprise has grown out of an extensive body of research on uncertainty. Yet, although widely acknowledged and studied, uncertainty remains a difficult concept to define or codify. Different conceptualizations and approaches to uncertainty abound in the literature, cross numerous fields of study, and touch a wide range of problem types. Two basic options are invariably followed in the face of uncertainty. The first is to reduce the uncertainties through data collection, research, modeling, simulation techniques, and so forth. Following this option, the objective is to overcome uncertainty, to make the unknown known. But the daunting nature of uncertainties surrounding global environmental change, as well as the need to make decisions before the normal science option can provide resolution, force a second option—that of managing or integrating uncertainty directly into the decision-making or policy-making process. Before uncertainty can be so integrated, however, the nature and extent of the uncertainty must be clarified.

The fields of mathematics, statistics, and, more recently, physics provide the "science of uncertainty" with many powerful means and techniques to conceptualize, quantify, and manage uncertainty, ranging from the frequency distributions of probability theory to the possibility and belief statements of Bayesian statistics. Addressing other aspects of uncertainty, fuzzy set logic offers an alternative to classical set theory for situations where the definitions of set membership are

vague, ambiguous, or nonexclusive. More recently, researchers have proposed chaos theory and complexification theory to focus on expecting the unexpected in models and theory (Casti, 1994).

The practical application of many of these techniques was originally pioneered by researchers in decision analysis (see Raiffa, 1968). In the fields of economics and decision theory, researchers continue to study rational decision making under uncertainty and how to assess the value of collecting additional information (Clemen, 1991). Methods for modeling risk attitudes, leading to the terms *risk-prone* and *risk-averse*, attempt to capture how different people faced with making a decision react to the uncertainty surrounding the expected outcomes. Uncertainty and its related context, surprise, are treated largely as the realization that events, currently unknown, will occur affecting the final outcome of a process or decision.

This acknowledgment of uncertainty has found a prominent place in many other fields of study, each one speaking its own language of uncertainty. Research on uncertainty cross-cuts a number of different disciplines. For example, researchers making risk assessments and setting safety standards find it most useful to distinguish between risk (the probability of a certain negative effect resulting from a hazard occurrence, given the specified level of exposure), variability (interindividual differences in vulnerability and susceptibility), and uncertainty (model parameter variability and any unexplained residual). In work related to hazards and risk, sociologists, anthropologists, psychologists, and geographers have made important contributions to the discussions on risk perception, risk communication, and the social amplification of risk (Kahneman et al., 1982; Kasperson et al., 1988; see Gigerenzer, 1996, for a criticism). Similarly, work on visualizing or graphically conveying uncertainty also crosses a diverse set of disciplines including psychology, computer science, and geographic information systems (GIS) (MacEachren, 1992).

Wynne (1992) emphasizes that the modeling of environmental risk systems requires examination of not only the scientific evidence and competing interpretations, but also investigation of the nature, assumptions, and inherent limitations of the scientific knowledge behind the data and the model. He specifies four types of uncertainty—risk, uncertainty, ignorance, and indeterminacy—each overlaying dimensions of uncertainty. *Risk* refers to a situation when the system behavior is well known and the chances of different outcomes can be quantified by probability distributions. If, however, the important system parameters are known but not the associated probabilities, then in Wynne's definitions, *uncertainty* exists. *Ignorance* is that which is not known (or even that we are aware that we do not know it) and, for Wynne, is endemic because scientific knowledge must set the bounds of uncertainty in order to function. *Indeterminacy* captures the unbounded complexity of causal chains and open networks. Uncertainty, in part, stems not only from an incomplete understanding of determinate relationships, but also from the interaction of these relationships with contingent and unpredictable actors and processes. However, the extent to which situations are truly "indeterminate," as opposed to simply containing a very broad distribution of subjective probability estimates, is not a straightforward classification, for often very ill understood phenomena can still be bounded to some extent by existing knowledge, and thus are not truly indeterminate.

3 OVERCOME OR JUST MANAGE UNCERTAINTY

In the areas of environmental policy and resource management, policymakers struggle with the need to make decisions utilizing vague and ambiguous concepts (such as sustainability), with sparse and imprecise information, in decisions that have far-reaching, and often irreversible, impacts on both environment and society. Not surprisingly, efforts to incorporate uncertainty into the decision-making process quickly move to the forefront with the advent of decision-making paradigms, such as the precautionary principle, adaptive environmental management, the preventative paradigm, or stewardship. Ravetz (1986) takes the concept of "usable knowledge in the context of incomplete science" one step further by introducing the idea of usable ignorance. To Ravetz, acknowledging the "ignorance factor" means becoming aware of the limits of our knowledge. Ravetz argues that ignorance cannot be overcome with any amount of sophisticated calculations. Rather, coping with ignorance demands a better articulation of the policy process and a greater awareness of how that process operates. He recognizes that one can only replace ignorance by gaining more knowledge, but stresses that by "being aware of our ignorance we do not encounter disastrous pitfalls in our supposedly secure knowledge or supposedly effective technique" (p. 429).

The emphasis on managing uncertainty rather than mastering it can be traced to work on resilience in ecology (Holling, 1986). Whereas resistance implies an ability to withstand change or impact within some measure of performance, resilience captures the ability to give with the forcing function, without disrupting the overall health of the system. In this framework, adaptation is an ecological mechanism whose aim is not to overcome or control environmental uncertainty but to live with and, in some cases, thrive upon it.

Risk is typically defined as the condition in which the event, process, or outcome, and the probability that each will occur, is known. In reality, of course, complete or perfect knowledge of complex systems, which would permit the credible calculation of objective or frequentist probabilities, rarely exists. Likewise, the full range of potential outcomes is usually not known. Thus, risk almost always is accompanied by varying degrees of uncertainty. Uncertainty is usually defined as the condition in which the event, process, or outcome is known (factually or hypothetically), but the probabilities that it will occur are not known or are highly subjective estimates (see, e.g., Moss and Schneider, 2000).

4 SURPRISE

Strictly speaking, surprise is the condition in which the event, process, or outcome is not known or expected. In this "strict" meaning, the attribution of surprise shifts toward the event, process, or outcome itself. We may expect surprises to occur, but we are surprised by the specific event, process, or outcome involved. This meaning, as noted, begs the issue of anticipation because the very act of anticipation implies some level of knowledge or foresight. However, it may be possible to identify

"imaginable conditions for surprise" where the conditions that might induce surprises are known even though the actual surprise events are not—e.g., rapid forcing of nonlinear systems (Moss and Schneider, 2000).

Because of the impracticality of the strict definition of surprise for policy making, various studies advocate the use of another meaning for surprise, one in which the attribution of surprise shifts more toward the expectations of the observer. Holling (1986; p. 294) recognized this meaning of surprise as a condition in which perceived reality departs qualitatively from expectations. It is this more interpretive or relational meaning of surprise—which has been labeled imaginable surprise—that portends to be most useful for global change studies [e.g., see Schneider et al. (1998) from which much of this material has been adopted].

Almost every event may constitute an imaginable surprise to someone. But since global change phenomena and their environmental and societal impacts are a community-scale set of issues, little can be gained for our purposes by focusing on whether someone, somewhere, may or may not have once predicted or hinted at some surprise event. More fruitful is the recognition that groups, communities, and cultures may share expectations such that a particular event is likely to qualify as a surprise for most within them. In these cases, what gets labeled as a surprise depends upon the extent to which reality departs from community expectations, and on the salience of the problems imposed.

Imaginable surprise applies to communities of experts, policymakers, managers, and educators who share common ranges of expectation that are generated by group dynamics, leaders, and signal processors, including the dominant educational and research paradigms (Kasperson et al., 1988). For these communities, shared expectations follow from dominant interpretations among the expert community (e.g., global warming is likely), from their fit with broader policy agendas (e.g., environmentally benign economic development is possible), and from vested interest, conscious or unconscious, of an agency or group to maintain a particular view (e.g., global population growth is environmentally damaging, or, alternatively, good for the economy). Since policy making often reflects a blend of public and interest group perceptions of reality, the imaginable surprise formulation is much more relevant to global change policy issues than a strict definition of surprise as an unimaginable outcome.

5 APPLICATION TO GLOBAL CHANGE

Since natural and social global change science remains in a range of developmental stages, the unknowns are sufficiently large to warrant attention to divergent themes about similar processes and outcomes. To facilitate this range of research, (a) measures should be taken to ensure a more open discourse and evaluation of alternatives, such as by a more open airing and professional evaluation as opposed to uncritical, "equal time," and equal credibility often afforded to polarized viewpoints in the popular media of less dominant or unconventional views, including those by advocacy science and scientists; and (b) by reducing the redundancy of research

focused on the dominant views and theses while still preserving a diversity of approaches within dominant paradigms—i.e., create research "overlap without cloning" (Schneider et al., 1998).

The assumptions associated with the standard paradigm of global climate change impact assessment, for example, although recognizing the wide range of uncertainty, are essentially surprise free. One approach is to postulate low, or uncertain, probability cases in which little climate change, on the one hand, or catastrophic surprises, on the other hand, might occur and multiply the lower probability times the much larger potential costs or benefits. Analysts, however, customarily use a few standard general circulation model CO_2-doubling scenarios to "bracket the uncertainty" rather than to postulate extremely serious or relatively negligible climatic change outcomes (e.g., see Schneider, 2001). A strategic approach, that is, one that considers a wide range of probabilities and outcomes, may be more appropriate for global climate change impact assessments given the high plausibility of surprises, even if we have but limited capacity to anticipate specific details right now (Moss and Schneider, 2000).

An assumption in cost–benefit calculations within the standard assessment paradigm is that "nature" is either constant or irrelevant. For example, ecological services such as pest control or waste recycling are assumed as constants or of no economic value in most assessment calculations. Yet should climatic change occur in the middle to upper range of that typically projected, it is highly likely that communities of species will be disassembled, and the probability of significant alterations to existing patterns of pests and weeds seem virtually certain (Root, 2000). Some argue that pests, should their patterns be altered, can simply be controlled by pesticides and herbicides. The side effects of many such controls are well known. What is not considered in the standard paradigm is the consideration of a "surprise" scenario such as a change in public consciousness regarding the value of nature that would reject pesticide or herbicide application as a "tech-fix" response to global changes.

Finally, global change portends alterations to the basic processes that govern the state of the biosphere. Global change research, therefore, might do well to anticipate these alterations, an effort that will require more than the study of extant processes and conditions alone. Various modes of analysis and approaches appropriate for such explorations, but typically underutilized in the research community, should be encouraged. Among these are (a) backcasting scenarios from posited future states and/or reconstructing past scenarios in alternative ways to identify events or processes that might happen (recognizing, of course, that diffusion processes usually are not reversible and diffusion-dominated systems cannot be uniquely backcast); (b) increasing attention to and support for the study of "outlier" outcomes, searching for the reasons they appear deviant and the lessons that might be drawn from them (Hassol and Katzenberger, 1997); and, (c) exploring the "resilience" paradigm (e.g., precautionary principle) alongside the "optimization" paradigm (e.g., aggregated cost–benefit analyses) to inform policy making and diagnose alternative outcomes and risk management strategies. Other means of improving the anticipation of surprise in global change science would emerge from convening additional expert

groups and asking them for more exhaustive assessments of the issues. Balanced assessments will likely lead to recommendations that "research as usual" be tempered with more alternative or even unusual research.

In summary, global change science and policy making will have to deal with uncertainty and surprise for the foreseeable future. Thus, more systematic analysis of surprise issues and more formal and consistent methods of incorporation of uncertainty into global change assessments will become increasingly necessary. Improvements in dealing with scientific surprise in climate change in particular and global change in general, therefore, require the research and funding communities to seek a better balance among traditional and experimental research alternatives (see also Kates and Clark, 1996, p. 31). This aim, in turn, requires strategies that will facilitate this balance, including the difficult problem of assessing "quality" in an interdisciplinary context.

ACKNOWLEDGMENTS

Modified from S. H. Schneider, B. L. Turner, and H. Morehouse Garriga (1998), "Imaginable Surprise in Global Change Science," *Journal of Risk Research* **1**(2): 165–185. I thank Kristin Kuntz-Duriseti for editorial help and Billie Lee Turner for the co-chairing the effort that produced much of the perspective here.

REFERENCES

Casti, J. L., *Complexification: Explaining a Paradoxical World through the Science of Surprise*, Harper Collins, New York, 1994.

Clemen, R. T., *Making Hard Decisions: An Introduction to Decision Analysis*, PWS-Kent, Boston, 1991.

Darmstadter, J., and M. A. Toman (Eds.), Nonlinearities and surprises in climate change: An introduction and overview, in *Assessing Surprise and Nonlinearities in Greenhouse Warming. Proceedings of an Interdisciplinary Workshop*, Resources for the Future, Washington, DC, 1993, pp. 1–10.

Gigerenzer, G., On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky, *Psychol. Rev.*, *103*, 592–596, 1996.

Hassol, S. J., and J. Katzenberger (Eds.), *Elements of Change 1996*, Aspen Global Change Institute, Aspen, CO, 1997.

Holling, C. S., The resilience of terrestrial ecosystems: Local surprise and global change, in W. C. Clark, and R. E. Munn (Eds.), *Sustainable Development of the Biosphere*, Cambridge University Press for the International Institute for Applied Systems Analysis, Cambridge, 1986, pp. 292–317.

Kahneman, D., P. Slovic, and A. Tversky, *Judgment under Uncertainty*, Cambridge University Press, New York, 1982.

Kasperson, R. E., O. Renn, P. Slovic, H. Brown, J. Emel, R. Goble, J. X. Kasperson, and S. J. Ratick, The social amplification of risk: A conceptual framework, *Risk Anal. 8*, 177–187, 1988.

Kates, R. W., and W. C. Clark, Environmental surprise: Expecting the unexpected? *Environment, 38*, 6–11, 28–34, 1996.

MacEachren, A., Visualizing uncertain information, *Cartogr. Perspect.*, (13), 10–19, 1992.

Moss, R. H., and S. H. Schneider, Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting, in R. Pachauri, T. Taniguchi, and K. Tanaka (Eds.), *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*, Intergovernmental Panel on Climate Change, Geneva, 2000; available on-line, http://www.gispri.or.jp.

Raiffa, H., *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Addison-Wesley, Reading, MA, 1968.

Ravetz, J. R., Usable knowledge, usable ignorance: Incomplete science with policy implications, in W. C. Clark and R. E. Munn (Eds.), *Sustainable Development of the Biosphere*, Cambridge University Press, New York, 1986, pp. 415–432.

Root, T. L., Ecology: Possible consequences of rapid global change, in G. Ernst (Ed.), *Earth Systems: Processes and Issues*, Cambridge University Press, Cambridge, MA, 2000, pp. 315–324.

Schneider, S. H., What constitutes "dangerous" climate change? *Nature, 411*, 17–19, 2001.

Schneider, S. H., B. L. Turner II, and H. Morehouse Garriga, Imaginable surprise in global change science, *J. Risk Res., 1*(2), 1998, 165–185.

Wynne, B., Uncertainty and environmental learning: Reconceiving science and policy in the preventive paradigm, *Global Environ. Change, 20*, 111–127, 1992.

Index Term **Links**

Index Term **Links**

Index Term	Links

Index Term **Links**