NONLINEAR DYNAMICS IN GEOSCIENCES

Edited by
Anastasios A. Tsonis
and
James B. Elsner

Aegean Conferences
Springer

Nonlinear Dynamics in Geosciences

Anastasios A. Tsonis, James B. Elsner
(Editors)

Nonlinear Dynamics in Geosciences

Springer

Anastasios A. Tsonis
Department of Mathematical Sciences
University of Wisconsin
Milwaukee, Wisconsin 53201
USA
aatsonis@uwm.edu

James B. Elsner
Department of Geography
Florida State University
114 Ferndale Drive
Tallahassee, Florida 32306
USA
jelsner@fsu.edu

Library of Congress Control Number: 2007929284

ISBN 978-0-387-34917-6 e-ISBN 978-0-387-34918-3

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Preface

From June 12-16, 2006 over eighty scientists from around the world gathered in Rhodes, Greece to attend the conference "20 Years of Nonlinear Dynamics in Geosciences". The editors of this book organized the conference. It was sponsored by Aegean Conferences and was endorsed by the American Meteorological Society and the European Geosciences Union.

In the past two decades many symposia and sessions in major general assemblies have focused on nonlinear dynamics. However, most of these meetings were specialized and dedicated to specific topics. The aim of this conference was to bring scientists from diverse fields in Geosciences together, to discuss how nonlinear approaches to specific problems are applied, and what major advances have been achieved.

This book is a series of talks presented at the conference. According to the design of the conference there were no parallel sessions and the presentation of the talks had no particular order. The idea was for everybody to be exposed to everybody else's scientific problems and to allow everyone to see how different disciplines approach common themes. This philosophy was applied in the organization of this book as well. The papers have no particular order.

The question we had to frequently answer was why this conference title?. We were fortunate to have the founder of chaos theory answer it for us in his letter addressing the participants. We could not have explained it better. Edward N. Lorenz's letter follows.

Anastasios A. Tsonis
James B. Elsner

Letter from Edward N. Lorenz

It is a great disappointment to me that I am not able to be with you, nor to see my many friends and acquaintances among you, nor to welcome you to this conference on 20 years of nonlinear dynamics while standing in front of you. I am happy to see from the program that nonlinear dynamics is not only alive and well but also thriving.

One of my first thoughts upon learning of the conference was "Why 20 years, rather than something closer to 200?" Nineteenth-century turbulence theorists were well aware of the nonlinear nature of the equations governing the phenomenon that they studied, and they knew that it was this nonlinearity that had prevented them from attaining their goals. More recently, some 60 years ago when I was a student, I was taught that the equations governing atmospheric flow were nonlinear, and that they defied a general solution. The popular method of handling them was linearization about a known particular solution, often a steady solution but in any case a highly specialized one. Often the results compared well with something that had been observed; in those days agreement within 20 or 30 per cent or sometimes within a factor of 2 was considered equivalent to verification. In other instances there was no agreement at all, in which case it was less likely that someone would announce the results, and there seemed to be no rule for determining in advance whether or not the procedure would work.

Things changed with the advent of the computer. Particular solutions of nonlinear systems could now be determined with fair precision. Disagreement with observations would more likely result from use of an inappropriate equation than from an inadequate means of solution.

All this happened well over 20 years ago. What distinguishes the past 20 years from perhaps 40 or more is what distinguishes actual nonlinear dynamics from simply solving nonlinear equations numerically. From studying more and more numerical solutions of equations of various degrees of complexity, we have come to know what these solutions do and do not look like, and this new knowledge has sometimes suggested principles that we have subsequently verified analytically. Often we can look at a time series and recognize it as typical of the output of a certain type of nonlinear system. We have learned, for example, that a time-dependent solution may oscillate in one manner for a long time, then switch to a quite different form of behavior, only to switch back again after another long interval - a phenomenon, sometimes observed in nature, that was once thought to require a change in the governing law. Of course, our new understanding did not suddenly arrive in 1986, but it is about 20 years since it came of age.

We now have such terms as strange attractor, multifractals, ensemble forecasting, and self-organized criticality that were either uncoined or generally unfamiliar to earth scientists in the middle 1980's, while old words, such as chaos, have acquired meanings that they did not formerly possess. In the next few days you will be hearing how our new grasp of nonlinear dynamics has led to ever more numerous and more diversified findings. I regret not being in the audience to hear about them with you.

Edward N. Lorenz

Contents

Contributors

Sergey G. Abaimov
University of California-Davis, Department of Geology, Davis, CA, USA

Fernando Angulo-Brown
Instituto Politécnico Nacional, Departamento de Física, Escuela Superior de Física y Matemáticas, México DF, México

M.A. Athanasiu
Democritus University of Thrace, Department of Electrical and Computer Engineering, Xanthi, Greece

Daniel N. Baker
University of Colorado, Laboratory for Atmospheric and Space Physics, Boulder, CO, USA

Lennart Bengtsson
University of Reading, Environmental Systems Science Centre, Reading, UK

Sylvie Zongo Brizard
ELICO Laboratory, FRE CNRS n°2816, University of Science and Technologies of Lille, Lille, France

Armin Bunde
Justus-Liebig-Universität Giessen, Institut für Theoretische Physik III, Giessen, Germany

Sandro Calmanti
ENEA, Environmental Department, Rome, Italy

Leila M.V. Carvalho
University of São Paulo, Department of Atmospheric Sciences, São Paulo, Brazil

Francisco Cervantes-De la Torre
Universidad Autónoma Metropolitana Azcapotzalco, Departamento de Sistemas, Area de Sistemas Computacionales, México DF, México

Grzegorz J. Ciach
University of Iowa, College of Engineering, IIHR-Hydroscience & Engineering, Iowa City, IA, USA

Antoni M. Correig
Universitat de Barcelona, Departament d'Astronomia i Meteorologia, Barcelona, Spain

Alessandro Dell'Aquila
ENEA, Environmental Department, Rome, Italy

Gregory S. Duane
National Center for Atmospheric Research, Boulder, CO, USA

Jan F. Eichner
Justus-Liebig-Universität Giessen, Institut für Theoretische Physik III, Giessen, Germany

Efi Foufoula-Georgiou
St. Anthony Falls Laboratory, National Center for Earth Surface Dynamics, Department of Civil Engineering, University of Minnesota, Minneapolis, MN, USA

Lizzie S.R. Froude
University of Reading, Environmental Systems Science Centre, Reading, UK

Alexander Gershunov
Scripps Institution of Oceanography, Climate Research Division, University of California-San Diego, San Diego, CA, USA

Aslak Grinsted
University of Oulu, Department of Geophysics, Oulu, Finland; and Arctic Centre, University of Lapland, Rovaniemi, Finland

Vijay K. Gupta
University of Colorado, Department of Civil, Environmental and Architectural Engineering, Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA

Shlomo Havlin
Bar-Ilan University, Minerva Center and Department of Physics, Ramat-Gan, Israel

Toshihiko Hirooka
Kyushu University, Department of Earth and Planetary Sciences, Fukuoka, Japan

Kevin I. Hodges
University of Reading, Environmental Systems Science Centre, Reading, UK

Tai-Yin Huang
The Pennsylvania State University, Physics Department, University Park, PA, USA

Yongxiang Huang
ELICO laboratory, FRE CNRS n°2816, University of Science and Technologies of Lille, Lille, France; *and*SIAMM, University of Shanghai, Shanghai, P.R. China

Tomoko Ichimaru
Kyushu University, Department of Earth and Planetary Sciences, Fukuoka, Japan

A.C. Iliopoulos
Democritus University of Thrace, Department of Electrical and Computer Engineering, Xanthi, Greece

Svetlana Jevrejeva
Proudman Oceanographic Laboratory, Liverpool, UK

Abigail Jiménez
University of Western Ontario, Department of Earth Sciences, Biological, and Geological Sciences, London, ON, Canada

Charles Jones
University of California-Santa Barbara, ICESS, Santa Barbara, CA, USA

Jan W. Kantelhardt
Justus-Liebig-Universität Giessen, Institut für Theoretische Physik III, Giessen, Germany; *and* Martin-Luther-Universität Halle-Wittenberg, Institut für Physik und Zentrum für Computational Nanoscience, Halle, Germany

Lakshmi Kantha
University of Colorado, Department of Aerospace Engineering, Boulder, CO, USA

A.D. Kirwan, Jr.
University of Delaware, College of Marine and Earth Studies, Newark, DE, USA

Alexander J. Klimas
National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, MD, USA

Tomasz J. Kozubowski
University of Nevada, Department of Mathematics and Statistics, Reno, NV, USA

Witold F. Krajewski
University of Iowa, College of Engineering, IIHR-Hydroscience & Engineering, Iowa City, IA, USA

Yuhji Kuroda
Meteorological Research Institute, Tsukuba, Japan

Bruce L. Lipphardt, Jr.
University of Delaware, College of Marine and Earth Studies, Newark, DE, USA

Yulu Liu
SIAMM, University of Shanghai, Shanghai, P.R. China

Shaun Lovejoy
McGill University, Department of Physics, Montréal, QC, Canada

Zhiming Lu
SIAMM, University of Shanghai, Shanghai, P.R. China

Valerio Lucarini
University of Camerino, Department of Mathematics and Computer Science, PASEF, Camerino, Italy; *and* University of Bologna, Department of Physics, ADGB – Atmospheric Dynamics Group Bologna, Bologna, Italy

Ramon Macià
Universitat Politècnica de Catalunya, Departament de Matemàtica aplicada II, Barcelona, Spain; *and* Institut d'Estudis Catalans, Laboratori d'Estudis Geofísics "Eduard Fontserè", Barcelona, Spain

David M. McClung
Department of Geography, University of British Columbia, Vancouver, BC, Canada

Juan Carlos Molinero
SIAMM, University of Shanghai, Shanghai, P.R. China

John C. Moore
University of Lapland, Arctic Centre, Rovaniemi, Finland

Hitoshi Mukougawa
Kyoto University, Disaster Prevention Research Institute, Kyoto, Japan

A. Brad Murray
Duke University, Nicholas School of the Environment and Earth Sciences and Center for Nonlinear and Complex Systems, Durham, NC, USA

Keith M. Nordstrom
Northwest Research Associates, Colorado Research Associates Division, Boulder, CO, USA

Anna K. Panorska
University of Nevada, Department of Mathematics and Statistics, Reno, NV, USA

Carlos G. Pavia-Miller
Universidad Autónoma Metropolitana Azcapotzalco, Departamento de Ciencias Básicas, Area de Física de Procesos Irreversibles, México DF, México; *and* Instituto Politécnico Nacional, Departamento de Física, Escuela Superior de Física y Matemáticas, México DF, México

Giorgos P. Pavlos
Democritus University of Thrace, Department of Electrical and Computer Engineering, Xanthi, Greece

Cécile Penland
NOAA/ESRL/Physical Sciences Division, Boulder, CO, USA

Andrew C. Poje
City University of New York, Department of Mathematics, Staten Island, NY, USA

Antonio M. Posadas
University of Almeria, Department of Applied Physics, Almeria, Spain

Antti Pulkkinen
UMBC/GEST at NASA/GSFC, Greenbelt, MD, USA

Alejandro Ramirez-Rojas
Universidad Autónoma Metropolitana Azcapotzalco, Departamento de Sistemas, Area de Sistemas Computacionales, México DF, México

Paul J. Roebber
University of Wisconsin-Milwaukee, Department of Mathematical Sciences, Atmospheric Sciences Group, Milwaukee, WI, USA

John B. Rundle
University of California-Davis, Department of Geology and Center for Computational Science and Engineering, Davis, CA, USA

Paulo M. Ruti
ENEA, Environmental Department, Rome, Italy

Victor B. Sapozhnikov
St. Anthony Falls Laboratory, National Center for Earth Surface Dynamics, Department of Civil Engineering, University of Minnesota, Minneapolis, MN, USA

Daniel Schertzer
CEREVE, Ecole Nationale des Ponts et Chaussées, Marne-la-Vallee, France

François G. Schmitt
ELICO laboratory, FRE CNRS n°2816, University of Science and Technologies of Lille, Lille, France

Robert Shcherbakov
University of California-Davis, Department of Geology and Center for Computational Science and Engineering, Davis, CA, USA

Antonio Speranza
University of Camerino, Department of Mathematics and Computer Science, PASEF, Camerino, Italy

De-Zheng Sun
CU/CIRES/Climate Diagnostics Center & NOAA/ESRL/Physical Science Division, Boulder, CO, USA

Kristy F. Tiampo
University of Western Ontario, Department of Earth Sciences, Biological, and Geological Sciences, London, ON, Canada

Joseph J. Tribbia
National Center for Atmospheric Research, Boulder, CO, USA

Brent M. Troutman
US Geological Survey, Denver Federal Center, Lakewood, CO, USA

Anastasios A.Tsonis
University of Wisconsin-Milwaukee, Department of Mathematical Sciences, Atmospheric Sciences Group, Milwaukee, WI, USA

Donald L. Turcotte
University of California-Davis, Department of Geology, Davis, CA, USA

Mercè Urquizu
Institut d'Estudis Catalans, Laboratori d'Estudis Geofísics "Eduard Fontserè", Barcelona, Spain

Dimitris Vassiliadis
ST at NASA/GSFC, Goddard Space Flight Center, Greenbelt, MD, USA; *and* George Mason University, Space Weather Lab, Fairfax, VA, USA

Josep Vila
Institut d'Estudis Catalans, Laboratori d'Estudis Geofísics "Eduard Fontserè", Barcelona, Spain; *and* Universitat de Barcelona, Departament d'Astronomia i Meteorologia, Barcelona, Spain

Gabriele Villarini
University of Iowa, College of Engineering, IIHR-Hydroscience & Engineering, Iowa City, IA, USA

Renato Vitolo
University of Camerino, Department of Mathematics and Computer Science, PASEF, Camerino, Italy

Harald Yndestad
Aalesund University College, Aalesund, Norway

Melissa Zweng
University of Delaware, College of Marine and Earth Studies, Newark, DE, USA

1 Introducing Networks in Climate Studies

Anastasios A. Tsonis

University of Wisconsin-Milwaukee, Department of Mathematical Sciences, Atmospheric Sciences Group, Milwaukee, WI 53201-0413 USA, aatsonis@uwm.edu

Abstract. Advances in understanding coupling in complex networks offer new ways of studying the collective behavior of interactive systems and already have yielded new insights in many areas of science. Here we present a overview of the first uses of network in climate studies.

1 Introduction and Basic Principles of Networks

A network is a system of interacting agents. In the literature an agent is called a node. The nodes in a network can be anything. For example, in the network of actors, the nodes are actors that are connected to other actors if they have appeared together in a movie. In a network of species the nodes are species that are connected to other species they interact with. In the network of scientists, the nodes are scientists that are connected to other scientists if they have collaborated. In the grand network of humans each node is an individual, which is connected to people he or she knows.

There are four basic types of networks.

a) Regular (ordered) networks. These networks are networks with a fixed number of nodes, each node having the same number of links connecting it in a specific way to a number of neighboring nodes (Fig. 1.1, left panel). If each node is linked to all other nodes in the network, then the network is a fully connected network.

b) Classical random networks. In these networks (Erdos and Renyi 1960) the nodes are connected at random (Fig. 1.1, right panel). In this case the degree distribution is a Poisson distribution (the degree distribution, p_k, gives the probability that a node in the network is connected to k other nodes). The problem with these networks is that they are not very stable. Removal of a number of nodes at random, may fracture the network to non-communicating parts.

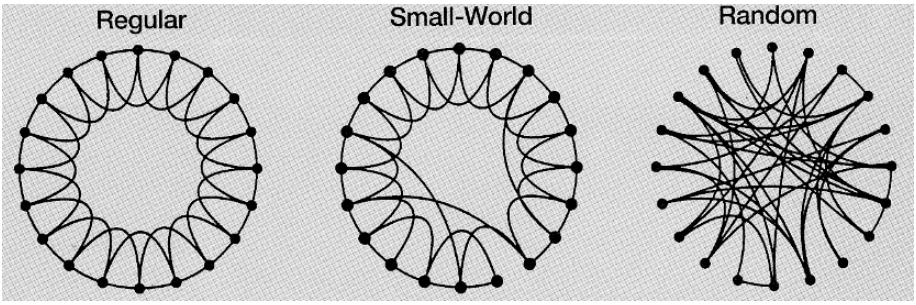


Fig. 1.1. Illustration of a regular, a small-world and a random network (after Watts and Strogatz, 1998, reproduced with permission from Science News).

c) Small-world networks. Regular networks are locally clustered, which means that, unless they are fully wired, it takes many steps to go from a node to another node away from its immediate neighborhood. On the contrary, random networks do not exhibit local clustering. Far away nodes can be connected as easily as nearby nodes. In this case information may be transported all over the network much more efficiently than in ordered networks. Thus, random networks exhibit efficient information transfer and regular networks do not (unless they are fully connected). This dichotomy of networks as either regular or random is undesirable since one could expect that in nature networks should be efficient in processing information and at the same time be stable. Work in this direction led to a new type of network, which was proposed a few years ago by the American mathematicians Duncan Watts and Steven Strogatz (1998) and is called *'small-world'* networks. A 'small-world' network is a superposition of regular and classical random graphs. Such networks exhibit a high degree of local clustering but a small number of long-range connections make them as efficient in transferring information as random networks. Those long-range connections do not have to be designed. A few long-range connections added at random will do the trick (Fig. 1.1, middle panel). Both random and 'small-world' networks are rather homogeneous networks in which each node has approximately the same number of links $<k>$. Both have nearly Poisson degree distributions that peak at $<k>$ and decay exponentially for large k.

d) Networks with a given degree distribution. The 'small-world' architecture can explain phenomena such as the six-degrees of separation (most people are friends with their immediate neighbors but we all have one or two friends a long way away), but it really is not a model found often in the real world. In the real world the architecture of a network is neither random nor 'small-world' but it comes in a variety of distributions such as truncated power-law distributions (Newmann 2001), gaussian distributions (Amaral, Scala, Barthelemy, and Stanley 2000), power-law distributions (Faloutsos, Faloutsos, and Faloutsos 1999), and distributions consisting of two

power-laws separated by a cutoff value of k (Dorogovtsev and Mendes 2001; Ferrer and Sole 2001). The last two types emerge in certain families of networks that grow in time (Dorogovtsev and Mendes 2001; Barabasi and Albert 1999).

The most interesting and common of such networks are the so-called *scale-free* networks, in which the degree distribution is the power law $p_k \sim k^{-\gamma}$. Consider a map showing an airline's routes (Fig. 1.2). This map has a few hubs connecting with many other points (super nodes) and many points connected to only a few other points, a property associated with power law distributions. Such a map is highly clustered, yet it allows motion from a point to another far away point with just a few connections.

As such, this network has the *property* of 'small-world' networks, but this property is not achieved by local clustering and a few random connections. It is achieved by having a few elements with large number of links and many elements having very few links. Thus, even though they share the same property, the architecture of scale-free networks is different than that of 'small-world' networks. Such inhomogeneous networks have been found to pervade biological, social, ecological, and economic systems, the internet, and other systems (Albert, Jeong, and Barabasi 1999; Jeong, Tombor, Albert, Oltvai, and Barabasi 2000; Liljeros, Edling, Amaral, Stanley, and Aberg 2001; Jeong, Mason, Barabasi, and Oltvai 2001; Pastor-Satorras and Vespignani 2001; Bouchaud and Mezard 2000; Farkas, Jeong, Vicsek, Barabasi, and Oltvai

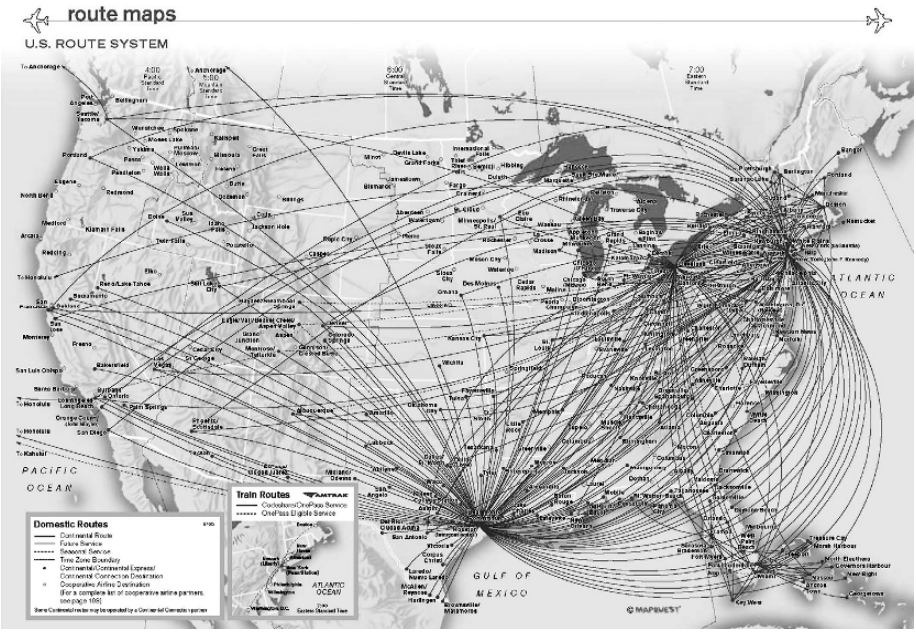


Fig. 1.2. Route map for Continental Airlines (courtesy of Continental Airlines).

2003; Barabasi and Bonabeau 2003; Albert and Barabasi 2002). These networks are referred to as scale-free because they show a power-law distribution of the number of links per node. Lately, it was also shown that, in addition to the power-law degree distribution, many real scale-free networks consist of self-repeating patterns on all length scales. This result is achieved by the application of a renormalization procedure that coarse-grains the system into boxes containing nodes within a given 'size' (Song, Havlin, and Maske 2005). In other words scale-free networks also exhibit fractal geometry. These properties are very important because they imply some kind of self-organization within the network. Scale-free networks are not only efficient in transferring information, but due to the high degree of local clustering they are also very stable (Barabasi and Bonabeau 2003). Because there are only a few super nodes, chances are that accidental removal of some nodes will not include the super nodes. In this case the network would not become disconnected. This is not the case with random and to a lesser degree with 'small-world' networks, where accidental removal of the same percentage of nodes makes them more prone to failure (Barabasi and Bonabeau 2003; Albert, Jeong, and Barabasi 2000). A scale-free network is vulnerable only when a super node is 'attacked'. Note that scale-free networks have properties of 'small-world' networks, but 'small-world' networks a la Watts and Strogatz are not scale-free. An example of such a network is given in Fig. 1.3, which shows the network of interactions between the proteins in the yeast *Saccharomyces cerevisiae*, otherwise known as baker's yeast (Jeong et al. 2001). By looking at the connectivity of each protein to other proteins, the authors were able to determine that

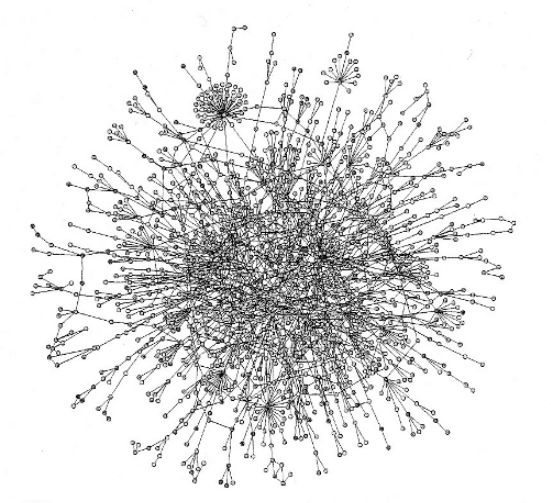


Fig. 1.3. The network of interactions between the proteins in the yeast Saccharomyces cerevisiae, otherwise known as baker's yeast (courtesy of A.-L. Barabasi).

more than 90% of the proteins in the network have less than five links and only one in five of these were essential to the survival of the yeast. In other words removing these proteins did not affect the function of this organism. In contrast, they found that less than 0.7% of the proteins were hubs having many more than fifteen connections. For these hubs they found that removal of any hub resulted in the death of the organism. Such findings, which can only be delineated by constructing the network, can be extremely useful as they may lead to ways to protect the organism from microbes by specifically protecting the hubs. In other areas, the presence of scale-free networks has led to strategies to slow the spread of diseases (Lijeros et al. 2001) and strategies to secure the internet (Barabasi and Bonabeau 2003).

The networks can be either fixed, where the number of nodes and links remains the same, or evolving, where the network grows as more nodes and links are added (some times in the literature growing networks are classified as a new type of network). Whatever the type of the network, its underlying topology provides clues about the collective dynamics of the network. The basic structural properties of networks are delineated by the clustering coefficient C and the characteristic path length (or diameter) L of the network. The clustering coefficient is defined as follows and is illustrated in Fig. 1.4: Assume that a node i is connected to k_i other nodes. Now con sider the k_i closest nodes of i. This defines the neighborhood of i. Then count the number of links, Δ_i, between any two nodes of the neighborhood (excluding node i). The clustering coefficient of node i is then given by $C_i=2\Delta_i/k_i(k_i-1)$. Since there can be at most $k_i(k_i-1)/2$ links between k_i nodes (which will happen if they formed a fully connected subnetwork), the clustering coefficient is normalized on the interval [0,1]. The average C_i over all nodes provides C. As such C provides a measure of local "cliqueness". The diameter of the network is defined by the number of connections in the shortest path between two nodes in the network averaged over all pairs of nodes. For a random network having the same average number of connections per node, $<k>$, it can be shown analytically (Bollabas 1985; Watts and Strogatz 1998; Albert and Barabasi 2001) that $L_{random}=lnN/ln<k>$ and $C_{random}=<k>/N$. The 'small-world' property requires that $C>> C_{\text{random}}$ and $L\geq L_{\text{random}}$. There are other measures and ways to investigate networks. Examples include minimum spanning trees (Mantegna 1998), asset trees and asset graphs (Onnela, Kaski, and Kertesz 2004), tree length and occupation levels (Onnela, Chakraborti, Kaski, Kertesz, and Kanto 2003), and intensity and coherence of networks (Onnela, Saramaki, Kertesz, and Kaski 2005) among others.

2 Examples of the Application of Networks to Climate

2.1 Climate networks

One way to delineate networks in the climate system is to assume that climate is represented by a grid of oscillators each one of them representing a dynamical sys tem varying in some complex way. What we are then interested in, is the collective

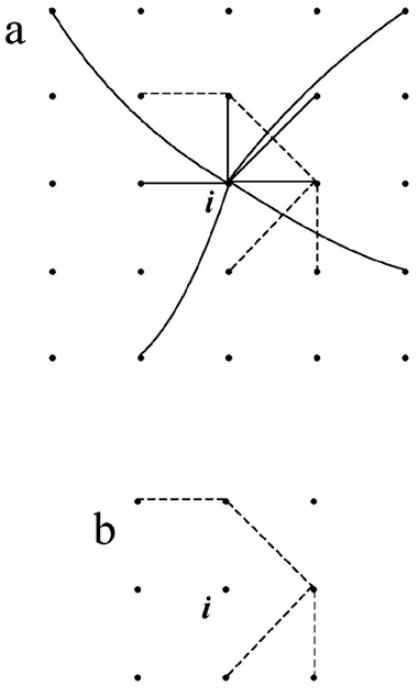


Fig. 1.4. Illustration of how to estimate the clustering coefficient. In (a) a node i is connected to k_i=8 other nodes (solid lines). In (b) we consider the k_i=8 closest nodes of i. This defines the neighborhood of i. In this neighborhood there exist Δ_i=4 connections between nodes (excluding node i) (broken lines). The clustering coefficient of node i is then given by $C_i=2\Delta_i/k_i(k_i-1)=0.143$. The average C_i over all nodes in the network provides the clustering coefficient of the network, C.

behavior of these interacting dynamical systems and the structure of the resulting network. Next, we will present an example, which will introduce us to the applications and promise of networks in atmospheric sciences. Some of these ideas have been presented in two recent publications by Tsonis and Roebber (2004) and Tsonis, Swanson, and Roebber (2006).

We start by considering the global National Center for Environmental Prediction/ National Center for Atmospheric Research (NCEP/NCAR) reanalysis 500 hPa data set (Kistler and coauthors 2001). A 500 hPa value indicates the height of the 500 hPa pressure level and provides a good representation of the general circulation (wind flow) of the atmosphere. The data used here are arranged on a grid with a resolution of 5^o latitude x 5^o longitude. For each grid point monthly values from 1950 to 2004 are available. This results in 72 points in the east-west direction and 37 points in the north-south direction for a total of n=2664 points. These 2664 points will be assumed to be the nodes of the network. From the monthly values we produced anomaly

values (actual value minus the climatological average for each month). Thus, for each grid point we have a time series of 660 anomaly values. In order to define the "connections" between the nodes, the correlation coefficient at lag zero (*r)* between the time series of all possible pairs of nodes [$n(n\text{-}1)/2$= 3,547,116 pairs] is estimated. Note that even though *r* is calculated at zero lag, a "connection" should not be thought of as "instantaneous". The fact that the values are monthly introduces a time scale of at least a month to each "connection". Even though most of the annual cycle is removed by producing anomaly values, some of it is still present as the amplitude of the anomalies is greater in the winter than in the summer. For this reason, in order to avoid spurious high values of *r*, only the values for December, January and February in each year were considered. It follows that the estimation of the correlation coefficient between any two time series is based on a sample size of 165. Note that since the values are monthly anomalies there is very little autocorrelation in the time series. A pair is considered as connected if their correlation $|r| \geq 0.5$. This criterion is based on parametric and non-parametric significance tests. According to the t-test with N=165, a value of r=0.5 is statistically significant above the 99% level. In addition, randomization experiments where the values of the time series of one node are scrambled and then are correlated to the unscrambled values of the time series of the other node indicate that a value of r=0.5 will not arise by chance. The use of the correlation coefficient to define links in networks is not new. Correlation coefficients have been used to successfully derive the topology of gene expression networks (Farkas et al. 2003; de la Fuente, Brazhnik, and Mendes 2002; Featherstone and Broadie 2002; Agrawal, 2002), and to study financial markets (Mantegna 1999). The choice of r=0.5 while it guarantees statistical significance is somewhat arbitrary. The effect of different correlation threshold is discussed in Tsonis and Roebber (2004). In any case, one may in fact consider all pairs as connected and study the so called weighted properties of the network where each link is assigned a weight proportional to its corresponding correlation coefficient (Onnela et al. 2003, 2004). For the scope of this paper, however, we will keep things simple and consider that a pair is connected if the correlation coefficient is above a threshold.

Once we have decided what constitutes a link, we are ready to look at the architecture of this network and how does it relate to dynamics. Figure 1.5 provides a first insight to this question. It shows the area weighted number of total links (connections) at each geographic location. More accurately it shows the fraction of the total global area that a point is connected to. This is a more appropriate way to show the architecture of the network because the network is a continuous network defined on a sphere, rather than a discrete network defined on a two-dimensional grid. Thus, if a node i is connected to N other nodes at λ_N latitudes then its area weighted connectivity, $\widetilde{C}_i$, is defined as

$$\widetilde{C}_i = \sum_{j=1}^{N} \cos\lambda_j \Delta A \Big/ \sum\nolimits_{over\ all\ \lambda\ and\ \phi} \cos\lambda \Delta A \qquad (1)$$

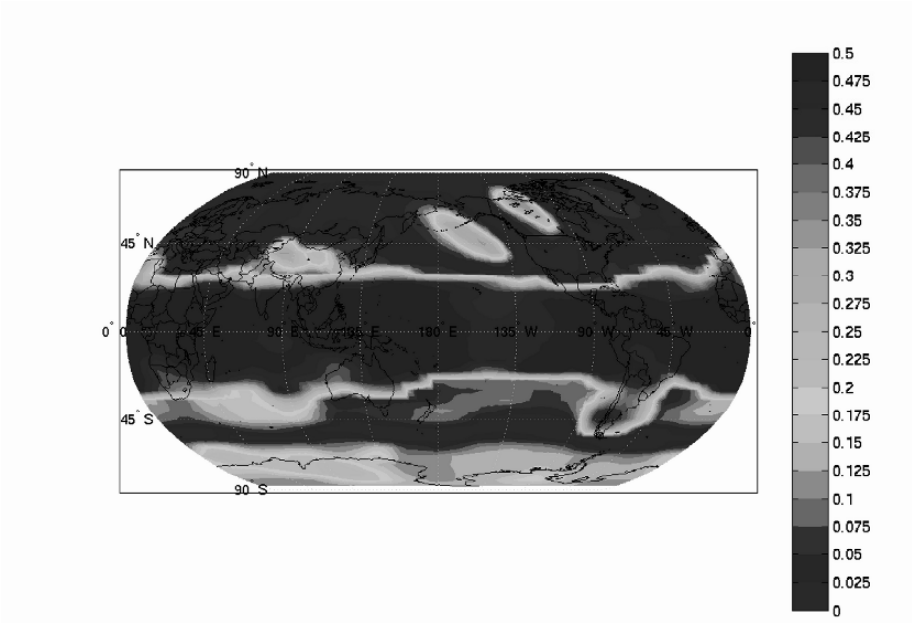


Fig. 1.5. Total number of links (connections) at each geographic location. The uniformity observed in the tropics indicates that each node possesses the same number of connections. This is not the case in the extratropics where certain nodes possess more links than the rest. See text for details on how this figure was produced. (A color version of this figure appears between pages 196 and 197).

where ΔA is the grid area and φ is the longitude. Once we have the information displayed in Fig. 1.5 we can estimate C and L. According to the definition of connectivity (equation 1), in order to find the clustering coefficient of node i, C_i, we consider a circular area on the sphere centered on i which is equal to $\widetilde{C}_i$. Then C_i is the fraction of this circular area that is connected (for a fully connected area, i.e. all pairs of nodes are connected, C_i=1 for all i). The average C_i over all nodes provides the clustering coefficient of the network, C. Note also that according to the definition of $\widetilde{C}$, the average $\widetilde{C}_i$ over all nodes gives the clustering coefficient C_{random}. Concerning the estimation of L, rather than finding the number of connections in the shortest path between two points, we estimate the distance of this path on the sphere. For this network we find that $L \cong$ 9,600 Km and C=0.56. For a random network with the same specifications (number of nodes, and average links per node) it is estimated that $L_{\text{random}} \cong$ 7,500 Km and C_{random}=0.19. These values indicate that indeed $L \geq L_{\text{random}}$ and $C > C_{\text{random}}$ (by a factor of three), which will make this global network close to a

'small-world' network. There is, however, more to this global network than what these values suggest.

Returning to Fig. 1.5 we observe that it displays two very interesting features. In the tropics it appears that all nodes posses more or less the same number of connections, which is a characteristic of fully connected networks. In the extratropics it appears that certain nodes posses more connections than the rest, which is a characteristic of scale-free networks. In the northern hemisphere we clearly see the presence of regions where such super nodes exist in China, North America and Northeast Pacific Ocean. Similarly several super nodes are visible in the southern hemisphere. These differences between tropics and extratropics are clearly delineated in the corresponding degree distributions. As is shown in Tsonis et al. 2006, the degree distribution for the network in Fig. 1.5 is a fusion of a scale-free network corresponding to the extratropics and an almost regular fully connected network corresponding to tropics. As is the case with all scale-free networks, the extratropical subnetwork is also a 'small-world' network. Indeed, we find that for points in the extratropics, the clustering coefficient is much greater than that of a corresponding random network (by a factor of nine). The collective behavior of the individual dynamical systems in the complete network is not described by a single type but it self-organizes into two coupled subnetworks, one regular almost fully connected operating in the tropics and one scale-free/'small-world' operating in the higher latitudes. The physical interpretation is that the equatorial network acts as an agent that connects the two hemispheres, thus allowing information to flow between them. This interpretation is consistent with the various suggested mechanisms for inter-hemispheric teleconnections (Tomas and Webster 1994; Love 1985; Compo, Kiladis, and Webster 1999; Meehl and coauthors 1996; Carrera 2001) and with the notion of sub-systems in climate proposed in the late 1980s (Tsonis and Elsner 1989; Lorenz 1991), and with recent studies on synchronized chaos in the climate system (Duane, Webster, and Weiss 1999).

An interesting observation in Fig. 1.5 is that super nodes may be associated with major teleconnection patterns. For example, the super nodes in North America and Northeast Pacific Ocean coincide with the well-known Pacific North America (PNA) pattern (Wallace and Gutzler 1981). In the southern hemisphere we also see super nodes over the southern tip of South America, Antarctica and South Indian Ocean that are consistent with some of the features of the Pacific South America (PSA) pattern (Mo and Higgins 1998). Interestingly, no such super nodes are evident where the other major pattern, the North Atlantic Oscillation (NAO) (Thompson and Wallace 1998; Pozo-Vazquez, Esteban-Para, Rodrigo, and Castro-Diez 2001; Huang, Higushi, and Shabbar 1998) is found. This does not indicate that NAO is not a significant feature of the climate system. Since NAO is not strongly connected to the tropics, the high connectivity of the tropics with other regions is masking NAO out. It should be noted here that in their pioneering paper Wallace and Gutzler (1981) defined teleconnectivity at each grid point as the strongest negative correlation between a grid point and all other points. This brings out teleconnection patterns associated with waves such as the trough-ridge-trough PNA pattern. However, because of the requirement of strongest negative correlation (which occurs between a negative anomaly center and a positive anomaly center), this approach can only delineate

long-range connections. As such, information about clustering and connectivity at other spatial scales is lost. In the network approach all the links at a point are considered and as such much more information (clustering coefficients, diameter, scaling properties etc) can be obtained. The similarities between Wallace and Gutzler's results and the network results arise from the fact that grid points with many long-range links will most likely stand out.

The physical interpretation of the results is that the climate system (as represented by the 500 hPa field) exhibits properties of stable networks and of networks where information is transferred efficiently. In the case of the climate system, "information" should be regarded as "fluctuations" from any source. These fluctuations will tend to destabilize the source region. For example, dynamical connections between the ocean and the atmosphere during an El Nino may make the climate over tropical Pacific less stable. However, the 'small-world' as well as the scale-free property of the extratropical network and the fully connected tropical network allow the system to respond quickly and coherently to fluctuations introduced into the system. This "information" transfer diffuses local fluctuations thereby reducing the possibility of prolonged local extremes and providing greater stability for the global climate system. An important consequence of this property is that local events may have global implications. The fact that the climate system may be inherently stable may not come as a surprise to some, but it is interesting that we find that a stable climate may require teleconnection patterns.

2.2 Detecting climate change

From this initial application of networks to climate it appears that atmospheric fields can be thought as a network of interacting points whose collective behavior may exhibit properties of 'small-world' networks. This ensures efficient transfer of information. In addition, the scale-free architectures guarantee stability. Furthermore, super nodes in the network identify teleconnection patterns. As was demonstrated in Tsonis (2004), these teleconnections are not static phenomena but their spatiotemporal variability is affected by large (global) changes. The 55-year period used to produce Fig. 1.5 can be divided into two distinct periods each of length of 27 years (1951-1977 and 1978-2004). During the first period the global temperature shows no significant overall trend. During the second period, however, a very strong positive trend is present. Since there is a distinct change in the global property of the system, how does this affect the dynamics of the global network? To answer this question C and L for the two periods were estimated. It was found that C is about 5% smaller and L is about 4% smaller in the second period. This result will indicate that during the warming of the planet the network has acquired more long-range connections and less small range-connections. This is clearly shown in Fig. 1.6, which shows the distribution of connections according to their distance (as calculated on the sphere). The solid line represents the distribution in the first period and the dashed line the distribution in the second period. This figure shows that the frequency of long-range connections (>6,000 km) has increased whereas the frequency

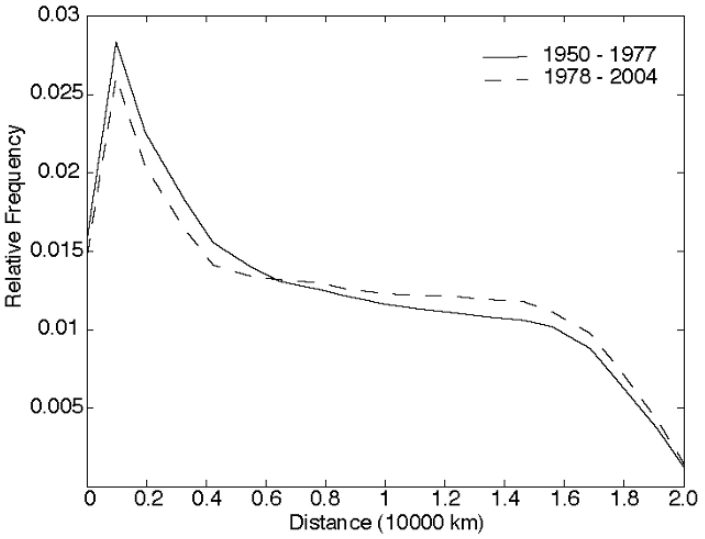


Fig. 1.6. The relative frequency distribution of the connections according to their distance for the period 1951-1977 (solid line) and for the period 1978-2004 (dashed line). This result indicates that during periods of warming, the network acquires more long-range connections and less small range-connections.

of shorter-range connections (<6,000 Km) has decreased. The differences between the two distributions may not appear impressive but with hundreds of thousands of connections involved, these differences are, according to the Kolmogorov-Smirnov test, statistically significant at the 99% confidence level (see also Tsonis 2004). Even though this is only one example, this result suggests that monitoring the properties of such networks may provide an additional tool to identify or verify climate change.

2.3 Atmospheric dynamics

Next we considered a network of major climate indices and assumed that each index represents a node of a network. The indices considered are the Pacific North America signal (PNA), the Pacific Decadal Oscillation (PDO), the North Atlantic Oscillations (NAO), the El Nino/Southern Oscillation (ENSO), the Western Pacific index (WP), and the Tropical North Atlantic signal (TNA). These six climate indices, which represent major climate modes, are assumed to be connected to each other with strength equal to the correlation coefficient between

them. A way to study their collective variation as a function of time is to estimate the mean distance *d(t)*

$$d(t) = \frac{2}{N(N-1)} \sum_{d_{ij}^t \in D^t} d_{ij}^t$$

where t denotes the time in the middle of a sliding window Δt, *N=6, i=1,N, l=1,N,* and $d_{ij}^t = \sqrt{2(1-|\rho_{ij}^t|)}$ and ρ_{ij}^t is the correlation coefficient between nodes *i* and *j* in the interval Δt. A distance of zero signifies a very well connected network and a distance of $\sqrt{2}$ signifies a set of non-connected nodes. Figure 1.7 shows the distance when all indices are included as nodes in this six-node network (black broken line). Also in this figure the distance when we exclude a single index is shown. For example, the yellow line shows the distance when the PDO is not part of the network (in other words when the network is a five-node network where the nodes are NAO, PNA, ENSO, TNA, and WP). The parallel thin line is the 95% significance level when the nodes are uncorrelated time series with the same lag1 autocorrelation as the

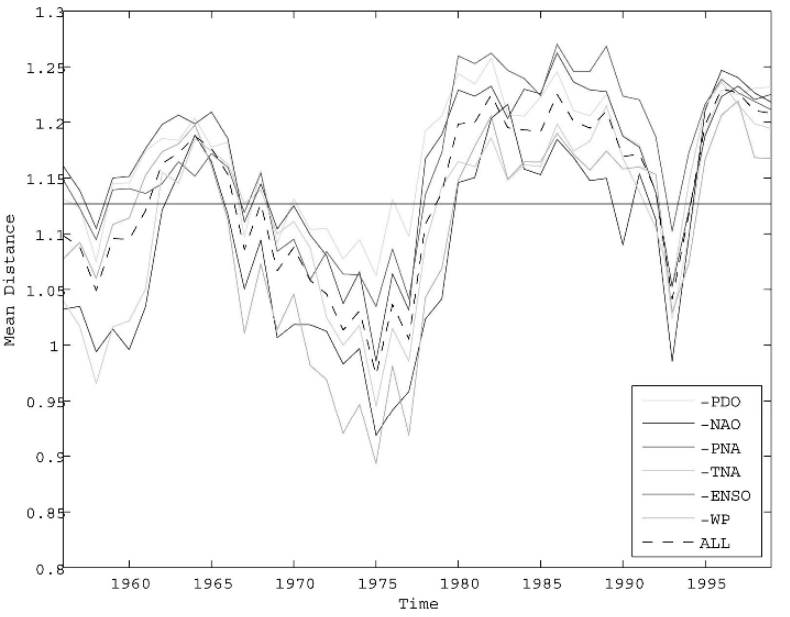


Fig. 1.7. The collective behavior of a network of six climate indices. See text for details. (A color version of this figure appears between pages 196 and 197).

climate indices. Anything outside this limit indicates that the network is real (i.e. the correlation coefficients between nodes are significantly different than zero). Subsequently, this figure indicates that overall the network is strongest when it does not include the NAO or the WP or the TNA (in these cases the distance of the five-node network is consistently smaller than the distance of the complete six-node network) and weaker when it does not include the PDO, or the ENSO or the PNA. This suggests that the NAO, WP, and TNA modes are dominated by climate noise, whereas the PDO, ENSO, and PNA arise primary from external forcing (dynamics). Research in this area is in progress and more results will be reported elsewhere.

3 Outlook

Complex network describe many natural and social dynamical systems and their study has revealed interesting mechanisms underlying their function. The novelty of networks is that they bring out topological/geometrical aspects which are related to the physics of the dynamical system in question thus providing a new and innovative way to treat and investigate nonlinear systems and data. While several advances have been made, this area is still young and the future is wide open. This introductory paper presented some fundamental aspects of networks and some preliminary results of the application of networks to climatic data, which indicate that networks delineate some key and new features of the climate system. This suggests that networks have the potential to become a new and useful tool in climate research.

Acknowledgements

This work was supported by NSF grant ATM-0438612

References

Albert, R., H. Jeong, and A-L Barabasi, 1999: Diameter of the World Wide Web. *Nature* 401, 130-131.

Albert, R., H. Jeong, and A.-L.Barabasi, 2000: Error and attack tolerance of complex networks. *Nature* 406, 378-382.

Albert, R., and A.-L. Barabasi, A-L. 2002: Statistical mechanics of complex networks. *Rev. Mod. Phys.*74, 47-101.

Amaral, L.A.N., A. Scala, M. Barthelemy, and H.E. Stanley, 2000: Classes of behavior of small world networks. *Proc. Natl. Acad. Sci. USA* 97, 11149-11152.

Agrawal, H., 2002: Extreme self-organization in networks constructed from gene expression data. *Phys. Rev. Lett.* 89, 268702.

Barabasi, A.-L., and E. Bonabeau, 2003: Scale-free networks. *Scientific American* 288, 60-69.

Bollabas, B., 2001: *Random Graphs*, 2nd Ed., Cambridge, 498 pp.

Bouchaud, J-P., and M. Mezard, 2000: Wealth condensation in a simple model of economy. *Physica A* 282, 536-540.

Carrera, M.L., 2001: Significant events of inter-hemispheric atmospheric mass exchange. Ph.D. thesis, McGill University, Montreal, Canada.

Compo, G.P., G.N. Kiladis, and P.J. Webster, 1999: The horizontal and vertical structure of east Asian winter monsoon pressure surges. *Quart J. Roy Meteor. Soc*. 125, 29-54.

Dorogovtsev, S.N., and J.F.F. Mendes, 2001: Language as an evolving word web. *Proc. Roy. Soc. London B 268*, 2602-2603.

Duane, G.S., P.J. Webster, and J.B. Weiss, 1999: Co-occurrence of northern and southern hemisphere blocks as partially synchronized chaos. *J. Atmos Sci.* 56, 4183-4205.

Erdos, P., and A. Renyi, 1960: On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* 5, 17-61.

Faloutsos, M., P. Faloutsos, and C. Faloutsos, 1999: On power-law relationships of the internet topology. *Comp. Comm. Rev.* 29, 251-260.

Farkas, I. J., H. Jeong, T. Vicsek, A.-L. Barabási, and Z.N. Oltvai, 2003: The topology of the transcription regulatory network in the yeast Saccharomyces cerevisiae. *Physica A* 318, 601-612.

Featherstone, D. E., and K. Broadie, 2002: Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *Bioessays* 24, 267-274.

Ferrer, R., and R.V. Sole, 2001: The small-world of human language. *Proc. Roy. Soc. London B* 268, 2261-2266.

de la Fuente, A., P. Brazhnik, and P. Mendes, 2002:Linking the genes: Inferring quantitative gene networks from microarray data. *Trends in Genetics* 18, 395-398.

Huang, J.P., K. Higuchi, and A. Shabbar, 1998: The relationship between the North Atlantic Oscillation and El Nino Southern Oscillation. *Geophys. Res. Lett.*. 25, 2707-2710.

Jeong, H., S. Mason, A.-L. Barabasi, and Z.N. Oltvai, 2001: Lethability and centrality in protein networks. *Nature* 411, 41-42.

Jeong, H., B. Tombor, R. Albert, A.N. Oltvai, and A.-L. Barabasi, 2000: The large scale organization of metabolic networks. *Nature* 407, 651-654,

Kistler, R., and Coauthors, 2001: The NCEP/NCAR 50-year reanalysis: monthly means, CD-ROM and documentation. *Bull. Amer. Meteor. Soc.* 82, 247-267.

Liljeros, F., C. Edling, L.N. Amaral, H.E. Stanley, and Y. Aberg, 2001: The web of human sexual contacts. *Nature* 411, 907-908.

Lorenz, E.N., 1991: Dimension of weather and climate attractors. *Nature* 353, 241-244.

Love, G., 1985: Cross-equatorial influence of winter hemisphere subtropical cold surges. *Mon. Wea. Rev.* 113, 1487-1498.

Mantegna, R.N., 1999: Hierarchical structure in financial markets. *Eur. Phys. J.* B 11, 193-197.

Meehl, G.A., and Coauthors 1996: Modulation of equatorial subseasonal convective episodes by tropical-extratropical interaction in the Indian and Pacific ocean regions. *J. Geophys. Res.* 101, 15,033-15,049.

Mo, K.C., and R.W. Higgins, 1998: The Pacific-South America modes and tropical convection during the southern hemisphere winter. *Mon. Wea. Rev*. 126, 1581-1596.

Newmann, M.E.J., 2001: The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci USA* 98, 404-409.

Onnela, J.-P., A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto, 2003: Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E* 68, 056110.

Onnela, J.-P., K. Kaski, and J. Kertesz, 2004: Clustering and information in correlation based financial networks. *Eur. Phys. J.* B 38, 353-362.

Onnela, J.-P., J. Saramaki, J. Kertesz, and K. Kaski, 2005: Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E*71, 065103.

Pastor-Satorras, R., and A. Vespignani, 2001: Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86, 3200-3203.

Pozo-Vazquez, D., M.J. Esteban-Parra, F.S. Rodrigo, and Y. Castro-Diez, 2001: The association between ENSO and winter atmospheric circulation and temperature in the North Atlantic region. *J Climate* 14, 3408-3420.

Song, C., S. Havlin, and H.A. Makse, 2005: Self-similarity of complex networks. *Nature* **433,** 392-395.

Strogatz, S.H., 2001: Exploring complex networks. *Nature* 410, 268-276.

Thompson, D.W.J., and J.M. Wallace, 1998: The Artic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett* 25(9), 1297-1300.

Tomas, R.A., and P.J. Webster, 1994: Horizontal and vertical structure of cross-equatorial wave propagation. *J. Atmos. Sci.* 51, 1417-1430.

Tsonis, A.A., 1992: *Chaos: From theory to applications*, Plenum Press, 274 pp.

Tsonis, A.A., and P.J. Roebber, 2004: The architecture of the climate network. *Physica A* 333, 497-504.

Tsonis, A.A., 2004: Does global warming inject randomness into the climate system? *EOS*, 85(38), 361-364.

Tsonis, A.A., K.L. Swanson, and P.J. Roebber, 2006: What do networks have to do with climate? *Bull. Amer. Meteor. Soc*. DOI:10.1175/BAMS-87-5-585.

Wallace, J.M., and D.S. Gutzler, 1981: Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Wea. Rev*. 109, 784-812.

Watts, D.J., and S.H. Strogatz, 1999: Collective dynamics of 'small-world' networks. *Nature* 393, 440-442.

2 Two Paradigms in Landscape Dynamics: Self-Similar Processes and Emergence

A. Brad Murray

Duke University, Nicholas School of the Environment and Earth Sciences and Center for Nonlinear and Complex Systems, abmurray@duke.edu

Abstract. Landscapes often exhibit self-similar (fractal, self-affine, multifractal) patterns, expressed in terms of power-law scalings. These scaling signatures are commonly interpreted as indicating that the same physical processes operate across a wide range of scales, as in the archetypical phenomenon of turbulence. Landscape researchers have also employed another productive perspective, that of emergent phenomena. To address relatively large-scale patterns and behaviors, rather than basing models directly on the vast number of small-scale processes occurring within the landscape, many models treat variables and interactions that arise from (and in turn influence) the collective behavior of smaller-scale degrees of freedom. The emergent-phenomena approach, in which processes at different scales are treated as different and only indirectly related, seems at first blush to be incompatible with the indications that landscape-forming processes can be scale invariant. However, analysis of research addressing landscape evolution suggests that the two perspectives can be complementary. Models that treat the coupled evolution of hillslopes and channels—synthesizing the effects of the many processes within these units—can explain the emergence of self-similarity (or self-affinity) of large-scale topography and river networks. Similarly, a model based on abstracted representations of the interactions between flow and topography offers an explanation for the basic scaling properties of braided streams. A model of the evolution of sandy-coastline shapes highlights how interactions between emergent structures can offer clear, direct explanations for scaling properties. In this case, the processes and patterns are scale-independent in the sense that the length scale changes over time, while a temporally self-similar shape persists, although at any time during the evolution a single wavelength dominates. The turbulence paradigm suggests that landscape systems, with many degrees of freedom, need to be treated as stochastic. However, models treating interactions between emergent landscape features show that complex, self-similar patterns and behaviors can be understood in terms of relatively simple, essentially deterministic processes.

1 Introduction: Conflicting Paradigms?

Various approaches and perspectives stemming from nonlinear dynamics and complex-systems research have percolated into the study of Earth-surface processes.

Here I will discuss two conceptual frameworks that have become paradigms: 1) what I will call the turbulence paradigm, in which essentially the same processes operate over a wide range of scales, giving rise to power-law scaling signatures (self-similarity, anisotropic self-affinity, fractals and multifractals—referred to here collectively simply as 'fractal' or 'self-similar'); and 2) the emergent phenomena viewpoint that processes at different scales differ in fundamental ways.

In the archetypical process of turbulence, the main story consists of a cascade through space and time scales, caused by a fascinating consistency of physical processes: eddies rubbing against each other and making smaller eddies, which then make even smaller vortices… Because the interactions between the scales are the same throughout, the statistics characterizing the activity at each scale vary in a consistent way as the scale of measurement shifts; a power-law scaling results. With myriad degrees of freedom—describing the vortices, for example—distributed smoothly across the scales, and myriad interactions between ever-changing combinations of vortices, deterministic models seem daunting, while stochastic descriptions are useful.

Power-law scaling, which also describes fractal and multi-fractal structures, has been discovered for many Earth-surface phenomena (Turcotte 1997), including topographic (Chase 1992; Gagnon et al. 2003; Mandelbrot 1983) and river-network (Gupta and Waymire 1989; Sapozhnikov and Foufoula-Georgiou 1996; Sapozhnikov and Foufoula-Georgiou; Sapozhnikov et al. 1998; Veitzer and Gupta 2000) patterns, and Earthquake and forest-fire (Turcotte et al. 2002) behaviors. Such a discovery suggests that, as in turbulence, a relatively simple set of scale-independent interactions could explain a sometimes dizzying range of scales—from river networks draining continents to annoying rills in a gravel driveway after a heavy rain, for example. Power-law scaling can also have direct practical implications for natural hazards, improving forecasts of the frequencies of rare, large-magnitude events (Rundle et al. 1996).

The emergent-phenomena concept implies in some sense the opposite view of the natural world—one in which, unlike the scale-free universality of a turbulence-like system, very different kinds of interactions can appear on different scales. Effectively new variables and interactions can arise from the collective behaviors of many degrees of freedom, the way fluid temperature, density, and pressure emerge from molecular motions. The emergent-phenomena viewpoint suggests that reductionism should be done in steps; we don't directly reduce a laboratory convection experiment to molecular dynamics, but explain the pattern formation through the interaction of macroscopic variables. In addition, the molecular motions are constrained by the emergent convection rolls; smaller-scale interactions are not necessarily more 'fundamental' than the ones defined on larger scales.

Thus, viewed at one scale a system can consist of stochastic interactions between a very large number of variables, while the same system can be understood on larger scales more simply through the analysis of a small number of variables—possibly even behaving as low-dimensional and deterministic. Restricting models (conceptual, analytical and numerical) to interactions between variables on scales not too much smaller than those of the phenomena of interest allows for more-direct mechanistic explanations (Murray 2003).

Earth scientists can certainly look at messy, spatially extended systems in nature and see many degrees of freedom, many complicated interactions, and 'noise'—a perspective that doesn't encourage the development of simple holistic models. However, with the key lesson from Chaos Theory that apparently random behavior can result from simple interactions (Lorentz 1963), and with an eye for emergent phenomena, a researcher can also look for a description in terms of relatively large-scale interactions that synthesize the collective effects of the many smaller-scale processes. Such a perspective has lead to many conceptual breakthroughs and associated numerical models for Earth-surface patterns and behaviors (Ashton et al. 2001; Chase 1992; Howard 1994; Kessler et al. 2001; Kessler and Werner 2003; Murray and Thieler 2004; Strudley et al. 2006; Werner and Fink 1993; Werner and Hallet 1993), including the examples presented in subsequent sections.

The emergent-phenomena approach suggests employing separate models for behaviors at different scales (Werner 1999; Werner 2003). Power-law scaling and the turbulence analogy, on the other hand, implies that one model should apply across a range of scales. Is there a conflict between these two highly productive paradigms, as this contrast suggests? An examination of research concerning terrestrial landscapes and river networks (Section 2) implies instead a complementary relationship, in which the origin of the fractal patterns are addressed using models of emergent interactions. Section 3 outlines a coastline-dynamics example in which emergent, large-scale interactions afford direct explanations of shoreline patterns and behaviors, including the evolution of a single dominant wavelength, which coexists with second-order multi-scale, power-law variations in shoreline shape.

2 Emergent Self-Similarity in Topography and River Patterns

The fractal structures of river networks and terrestrial topography are often related; at least in many settings they are likely caused by a single process operating over a broad range of scales: fluvial erosion (Chase 1992). In an erosional environment, streams and rivers cut down into the bedrock along their courses, tending to steepen the adjacent hillslopes. This induces various forms of gravity driven transport processes that tend to reduce the slopes, including soil 'creep' and landsliding (Ritter et al. 2001). The dominant sediment-transport processes on hillslopes (possibly including cliffs) depend on the combinations of climate, rock type, and river-incision rate. But the basic result is the same in the long term: hillslopes and the ridges or divides between eroding channels tend to evolve an equilibrium form that is coupled to the lowering of the channels (Reneau and Dietrich). The heights of ridges are related to the lateral distance between the channels they separate. Thus, if the network of branching channels has a self-similar structure, the topography will also scale (although generally in an anisotropic, self-affine way, because the heights of the divides are nonlinearly related to the distance between channels).

For such settings dominated by fluvial erosion, we must then ask how the fractal nature of the fluvial network develops. However, we cannot address the development of the network without considering the evolution of the topography, which determines

flow patterns. In a thought experiment, if we start with some (perhaps random) initial topography, the flow routing determines the flow strength in each segment of an incipient fluvial network. These flow strengths (measured for example by the 'stream-power index' of discharge multiplied by channel slope) affect incision rates (although those rates are also affected by rock type and other factors listed below). As channels begin to erode, hillslope processes alter the topography across the landscape. The changing topography changes the flow routing, and a coupled flow/topography evolution proceeds.

Given this general picture, different ways to view and study the system are possible. On the one hand, many exceedingly complex physical, biological, and chemical processes operate within each of the components of the landscape, and these can each be studied individually. Hillslope processes include: chemical and physical weathering that transforms rocks (of various lithologies) into mobile sediment (Anderson 2002); groundwater flow that affects weathering and landsliding and transports precipitation into the channels (Miall 1992); physical and biological transport mechanisms ranging from rainsplash to animal burrowing (Kyungsoo et al. 2005). Vegetation strongly affects each of these processes. Sediment transport and erosion in a bedrock river channel is at least as complicated when viewed at small scales, affected by processes including: hydrodynamics at the grain scale (Schmeeckle and Nelson 2003); interactions between different grains (Schmeeckle et al. 2001) and between grains and the bedrock, which cause bedrock abrasion; inhomogeneous and dynamic collections of sediment that cover and protect the bed (Sklar and Dietrich 2001); 'quarrying' of large chunks of bedrock from the river bed (Whipple et al. 2000); and the shape of the river channel (Stark 2006), which can be influenced by vegetation (Murray and Paola 2003). Each of these processes, and many more that influence the evolution of the landscape, are interesting in their own right, and studying them is necessary for understanding physical and ecological processes on their respective time and space scales. With a direct reductionist approach to studying the overall landscape shape, all of these processes would also need to be well understood and characterized—and then ultimately coupled together—to reproduce large-scale, long-term landscape evolution.

On the other hand, an emergent-phenomena perspective suggests that there may be a more effective way to address long-term landscape evolution. Rather than basing models directly on processes operating on much smaller and faster scales, we might try to synthesize the collective effects those processes have on larger space and time scales. For example, rather than resolving all the fascinating and dynamic processes within a bedrock channel, the main long-term effect they have—erosion—could be represented directly. Regardless of the small-scale stochastic details, in the long term the erosion rate must increase with discharge and with slope. (Or if the channel flows across previously deposited sediment, the sediment flux must increase with discharge and slope.) Similarly, regardless of the way precipitation is partitioned between ground water and surface flow during and shortly after an individual storm, in the long term precipitation will all be steered downhill, into and through the channel network, suggesting simplified flow-routing schemes. Hillslope processes tend to move more sediment down steeper slopes, and the essence of this collective

effect could be captured by a topographic-diffusion approach (sediment flux proportional to slope, or nonlinear equivalents (Yoo et al. 2005)).

Chase (1992) and Howard (1994) pioneered the development of models based on such emergent interactions. The large-scale parameterizations involved represent emergent phenomena in the same sense that macroscopic fluid properties and their interactions do. The scale differences between sediment grains and whole channels, or between soil pore spaces and hillslopes, are not as great as the scale differences between molecules and macroscopic fluid flow, but the approach of treating interactions between resulting relatively large-scale variables is the same. In addition, the initial large-scale parameterizations of landscape processes are not likely to be as quantitatively accurate as the Navier-Stokes equation, although those parameterizations tend to become more accurate as new measurements and the lessons of smaller-scale studies are synthesized. The works of Chase and Howard, and many subsequent models (Whipple and Tucker 1999; Willgoose et al. 1991), offer clearer insights into how the topography and drainage network co-evolve than would a hypothetical model based directly on the smaller- and faster-scale individual processes. Although such a hypothetical model would portray a very high-dimensional (many variables), stochastic system, the emergent interactions in the simpler models can be essentially deterministic (with random initial conditions typically). The emergent-interaction models also offer clearer potential explanations of landscape self-similarity. In detail, channels and associated processes look different in boulder-filled headwater streams (Furbish 1993) and larger sandy or gravely rivers (Parker 1978). However, the abstracted treatments in the simple models suggest that the long-term effects of channel processes (erosion related to discharge and slope) are similar across the scales. Below some (climate and rock-type dependent) spatial scale, topographic diffusion will fill in channels faster than stream processes can dig them. Above this scale, in a generally eroding region, channels all act to lower the landscape locally. Larger streams tend to lower the landscape more rapidly, so that smaller streams end up flowing into them, eventually creating a self-similar branching network.

Chase in particular treated the scaling aspects of the landscape as an emergent phenomenon to be explained in his paper “Fluvial landsculpting and the fractal dimension of topography” (Chase 1992). Since then, the scaling properties of landscapes have often been used as tests of model realism (Rodriguez-Iturbe et al. 1994; Rogriguez-Iturbe and Rinaldo 1997). And as multifractal methods of characterizing landscapes develop further, such synergistic interplay between the turbulence/scaling and emergent-phenomena paradigms holds continuing promise.

A similarly complementary combination of the paradigms occurred in braided-stream studies (Fig. 2.1). In some river reaches, the flow divides into numerous individual channels that perpetually bifurcate and rejoin each other in a lovely but dauntingly complicated dynamic pattern. Before the 1990’s, individual processes in braided rivers were studied largely in isolation. Random events such as boulders or logs lodging against each other were thought to be important in causing flow bifurcations (Leopold and Wollman 1957), but no mechanistic explanations for overall braiding dynamics were considered. Quantitative characterizations of the patterns

were limited to statistics such as the average number of channels in a cross-river transect.

Fig. 2.1. Braided rivers in the New Zealand Alps. (A color version of this figure appears between pages 196 and 197).

However, almost simultaneously the scaling and emergent-phenomena approaches were brought to bear on the decades-old enigma. Braiding patterns were shown to scale in space and time (Sapozhnikov and Foufoula-Georgiou 1996; Sapozhnikov and Foufoula-Georgiou 1999). And a model treating flow and sediment-transport interactions on scales commensurate with individual channels offered a holistic explanation in terms of fundamental feedbacks (Murray and Paola 1994; Murray and Paola 1997). (Because of the nonlinear relationship between flow strength and sediment flux, where the flow is converging and flow strength is increasing downstream, sediment flux must be diverging; erosion results where the flow gathers into a low area, causing the flow to converge more strongly. Similarly, where the flow is diverging downstream, deposition results. Deposition causes the flow to spread more, often resulting in channel division around a new bar.) These feedbacks operate locally, but interactions between the different parts of the spatially extended system lead to the self-organization of a perpetually dynamic, braided system. Although the output of the model, as for natural braiding, could only be characterized as stochastic, the interactions that caused it were simple and deterministic (random influences only appearing in the initial conditions)—a form of spatial-temporal chaos. The scale-free interactions in this model also offered an explanation for the existence of emergent self-similarity (self-affinity), and the scaling properties of natural and laboratory braided streams were used as a test of the model (Sapozhnikov et al. 1998). (Subsequently the large-scale approach to water and sediment routing has been improved and adopted in models addressing larger-scale, longer-term river-valley and drainage-basin evolution (Coulthard et al. 2002; Thomas and Nicholas 2002).)

3 Sandy-Coastline Shapes

As for terrestrial landscapes, different approaches are possible for analyzing how large-scale, plan-view shapes of sandy coastlines (e.g Fig. 2.2) evolve. Unlike coasts where prominent exposures of slowly eroding bedrock dominate the shoreline shape (typically steep, mountainous coasts), morphological change on largely sediment covered coastlines occurs through the rearrangement of sediment. In areas where tidal currents are relatively weak, waves and wave-driven currents are the chief sediment-moving agents in nearshore environments. From the smallest scales up, sediment transport starts with interactions on the time and space scales of waves sand grains, interactions that produce often-beautiful wave ripples (Fig. 2.3A) and other small-scale bedforms. These bedforms, captivating in their own right, also act as roughness elements, strongly affecting larger-scale patterns of current and sediment transport. The patterns of larger-scale flow, in turn, build an often complicated array of alongshore bars and channels in the region of breaking waves (the ‘surf zone’), as well as dents in the shoreline with similar scales (order of 100 m typically) (Fig. 2.3B). These bars and channels arise from, but also steer various kinds of surf-zone currents, including alongshore currents, and rip currents in the cross-shore channels.

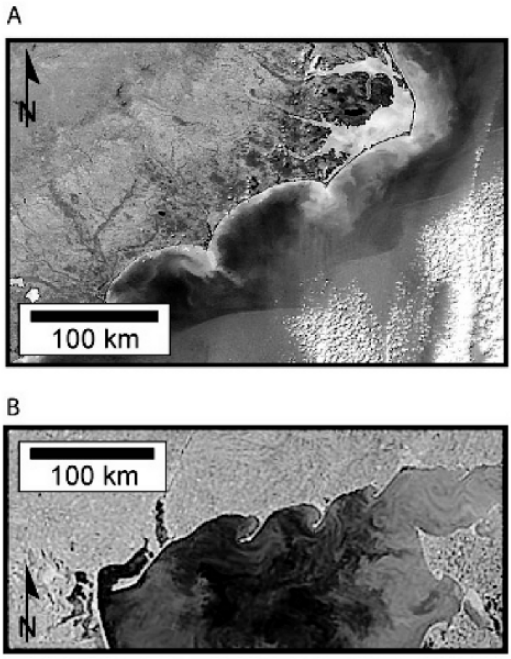


Fig. 2.2. Large-scale coastline shapes. A: The capes along the North and South Carolina coastline, USA. B: Flying spits in the Sea of Azov, Ukraine.

In the long term (> years), large-scale evolution of wave-dominated coastlines results largely from variations in alongshore sediment transport (Komar 1998). On a hypothetical smooth, straight shoreline, alongshore transport would be relatively easy to analyze. Breaking-wave momentum (evident to waders who are knocked shoreward in the surf zone) is deposited in the water column as the waves dwindle toward the beach. This momentum tends to drive the water in the direction of wave propagation, and because waves almost never move exactly toward shore, the alongshore component of this forcing generates an alongshore current. Although this current is usually subtle, it can transport considerable quantities of sediment because all the sand stirred up in the water by breaking-wave turbulence is advected alongshore. Alongshore gradients in the net alongshore sediment flux—averaged over all the wave directions and heights affecting a shoreline—can alter the shoreline shape. For example, moving along a shoreline in the direction of net sediment transport, if the sediment flux decreases alongshore, sediment accumulates, and the shoreline builds seaward in the long term (and vice versa). Gradients in alongshore sediment flux occur for various reasons, some of which are described below. (In the shorter term, during storms the shoreline erodes as sediment shifts from the beach to somewhat deeper parts of the nearshore profile, but these changes are essentially reversed (List and Ferris 1999) as the sand moves onshore again.)

A B

Fig. 2.3. Morphological patterns arising from different but coupled nearshore processes at different scales. A: Wave generated ripples (wavelength approximately 10 cm). B: Alongshore bars and channels shown by the wave breaking and gaps in the breaking, respectively, near Cape Hatteras, North Carolina, USA.

Alongshore currents and sediment transport are much more difficult to analyze in detail on realistic shorelines that are neither smooth nor straight. They arise somehow as the residual of all the complicated alongshore and cross-shore currents caused by the complicated pattern of waves breaking over bars, channels, and the beach (e.g. Fig. 2.3B). From this perspective, understanding and characterizing surf-zone dynamics seems to be a pre-requisite for addressing large-scale shoreline change. The bulk of nearshore-process research focuses on surf-zone-scale (or smaller scale) dynamics. While these interactions are fascinating by themselves, there may also be an implicit assumption that this is the best route to an eventual understanding of, and ability to predict, larger-scale, longer-term shoreline changes that are both scientifically compelling and of paramount concern for human coastal development.

However, surf-zone bars and channels come and go and shift around on times scales of days to weeks and typically only extend for tens to hundreds of meters. An emergent-phenomena perspective suggests that this may not be the most appropriate scale of processes on which to build models of coastline evolution over years to centuries that produces features spanning tens to hundreds of kilometers (e.g. Fig. 2.2). Stepping back from the transient details of breaking-wave patterns, we can draw some simple conclusions about large-scale wave-dominated coastline evolution.

3.1 An instability in coastline shape

The following thought experiment shows that alongshore sediment flux must be related to the relative angle between wave crests in deep water (before nearshore

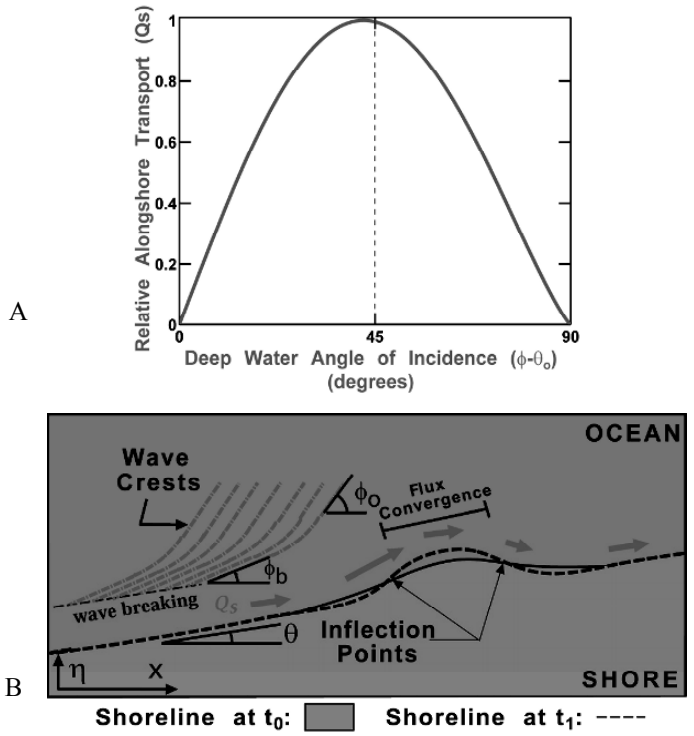


Fig. 2.4. An instability in plan-view coastline shape. A: A maximum in the relationship between alongshore sediment flux and deep-water wave-approach angles. B: The patterns of erosion and deposition that cause perturbations to a straight coastline to grow when waves approach from angles greater than the flux-maximizing angle ('high-angle' waves).

refraction and shoaling) and the shoreline orientation (averaged over spatial scales much greater than the alongshore extent of surf-zone bars): If deep-water wave crests are parallel to shore (a relative angle of 0°), there will be no alongshore component to the forcing where waves break, so alongshore flux (averaged alongshore over many bars) must be zero. If wave crests are perpendicular to shore (a relative angle of 90°), wave momentum and energy is not propagating toward shore, and again the alongshore sediment flux must vanish. Increasing the relative angle from 0° or decreasing it from 90° must both increase the sediment flux. So, the flux must reach a maximum for some angle between 0° and 90° (Fig. 2.4A).

This simple relationship has implications for coastline evolution. If deep-water waves approach from an angle greater than the one that maximizes the alongshore sediment flux ('high-angle' waves), a straight shoreline will be unstable—any deviations from a straight line will be exaggerated by such waves (Ashton et al. 2001) (Fig. 2.4B). Consider the convex-seaward portion of a subtle shoreline bump:

Moving along the shoreline there, the relative angle between deep-water crests and the local shoreline increases, moving farther from the angle that maximizes the alongshore flux. Therefore that flux must be converging, and the sediment accumulation will tend to cause the perturbation to grow seaward. A similar analysis shows that whenever deep-water waves approach from angles less than the flux-maximizing angle ('low-angle' waves), a bump will smooth out. For subtle shoreline perturbations, shoreline evolution can be described as diffusion of shoreline shapes for low-angle waves, and as anti-diffusion (negative diffusivity) for high-angle waves (Ashton et al. 2001; Falques 2003).

Whether a coastline becomes smoother or rougher in the long term depends on the relative influences of high- and low-angle waves. If the value of the flux-maximizing angle is known, analysis of historical wave records (e.g. from deep-water buoys) can yield a net diffusivity effective over the long term (Ashton et al. 2003a; Ashton and Murray 2006a; Ashton et al. 2003b). Such an analysis requires quantification of the relationship between alongshore sediment flux and deep-water wave approach angles.

The basis for such quantification is implicit in a commonly used relationship between alongshore sediment flux, Q_s, and breaking wave characteristics:

$$Q_s = K_1 H_b^{\frac{5}{2}} \cos(\phi_b - \theta) \sin(\phi_b - \theta), \quad (1)$$

where ϕ_b is the breaking wave crest angle θ is the shoreline orientation (Fig. 2.4B), and K_1 is an empirical constant. This relationship has been used in many coastal engineering models simulating shoreline evolution on scales up to years and kilometers. Because the height, H_b, and angle, ϕ_b, of breaking waves are themselves dependent on shoreline orientation, θ (which affects the extent of nearshore shoaling and refraction), the breaking wave description does not lead readily to the simple insights concerning shoreline evolution that the large-scale, deep-water-wave considerations do (Ashton et al. 2001; Murray and Ashton 2004). However, with the simplifying assumption of refraction and shoaling over shore-parallel bed contours, appropriate in the limit of large alongshore spatial scales (and short-period waves), Eq. 1 can be transformed into (Ashton et al. 2001):

$$Q_s = K_2 H_0^{\frac{12}{5}} T^{\frac{1}{5}} \cos^{\frac{6}{5}}(\phi_0 - \theta) \sin(\phi_0 - \theta), \quad (2)$$

where ϕ_b is the deep-water wave angle (Fig. 2.4B), and H_0 is deep-water wave height.

This relationship predicts that alongshore sediment flux is maximized for deep-water wave-approach angles of approximately 45° (consistent with other parameterizations of this relationship (Ashton and Murray 2006b)). Deep-water waves with relative angles greater than 45° are common, indicating that high-angle wave influences should be important on many coastlines. Where high-angle waves dominate coastline evolution, perturbations will grow over time. Along a spatially extended coastline, growing bumps of some sort could grow to finite amplitude and interact with each other. Exploring the modes of coastline evolution that could result requires a numerical model.

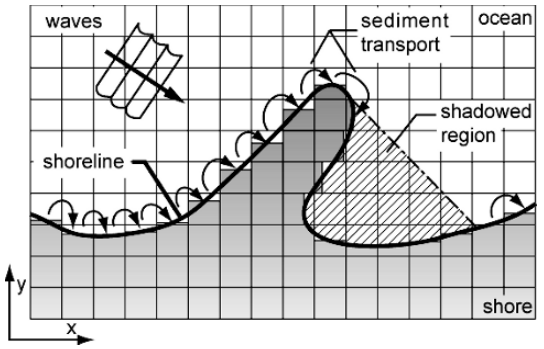


Fig. 2.5. Schematic illustration of the processes represented in the numerical model, including alongshore sediment flux (which depends on the local shoreline orientation) and wave shadowing (which prevents sediment transport in the lee of a coastline protuberance).

3.2 A simple model of large-scale coastline evolution

Ashton et al. (2001) constructed a simple numerical model based on alongshore sediment fluxes predicted by Eq. 1. Deep-water wave angle is input, and refraction and shoaling are assumed to occur over shore-parallel contours when determining local breaking-wave heights and angles. Alongshore gradients in alongshore sediment flux drive shoreline changes. Assuming that in the long term the cross-shore shape of the nearshore profile remains constant as it shifts landward or seaward (implicitly averaging out the shore-term fluctuations in profile shape related to storms), the conservation of nearshore sediment can be expressed as:

$$\frac{d\eta}{dt} = -\frac{1}{D}\frac{dQ_s}{dx}, \tag{3}$$

where η is the cross-shore shoreline position, D is the depth over which erosion or accretion are spread, and x is the alongshore coordinate. The model discretizes Eqs. 1 and 3, treating scales larger than those of nearshore bars.

As shoreline features protrude in the cross-shore direction, they can block the waves that would otherwise reach some other coastline segments. In such 'shadowed' regions, sediment transport does not occur in the model (Fig. 2.5).

A new deep-water wave direction is selected randomly from a probability distribution function representing a regional wave climate. As long as the distribution is weighted toward high-angle waves, interesting coastline behaviors result. Different kinds of features and interactions emerge under different wave climates, offering potential explanations for previously enigmatic phenomena including alongshore sandwaves (migrating plan-view shoreline undulations), flying spits, and cuspate capes (Fig. 2.6). The growth of cuspate capes in the model illustrates some of the emergent interactions responsible for the evolution of coastline shape.

Under a symmetrical or nearly symmetrical wave climate, where nearly as many waves approach from left and right looking offshore, symmetrical or nearly symmetrical capes and bays form (Fig. 2.6). At any time, all the capes tend to be approximately the same size, because smaller features grow faster in an anti-diffusional process (time scales are proportional to the square of the length scales in diffusion generally). However, the features also interact with each other in interesting ways. Because the capes are never precisely the same size, a slightly larger neighbor will always shadow a smaller neighbor from the highest-angle waves—which are the ones that produce the strongest anti-diffusion (Ashton and Murray 2006b). Thus, the larger neighbor grows faster, and the shadowing increases (Fig. 2.6A). At some point, so many of the high-angle waves are removed from the local wave climate felt at the tip of the smaller neighbor that the local wave climate becomes dominated by low-angle waves. When that happens, the smaller cape diffuses away rapidly. The progressive loss of capes corresponds to a ratcheting-up of the wavelength of the pattern in the long term (Fig. 2.6A). Interestingly, given the spatially varying effective shoreline diffusivities (including positive and negative diffusivities) and nonlocal interactions, the wavelength increase approximately follows a diffusional time scaling (Fig. 2.7). With enough time, model capes attain a wavelength as large as that of the largest natural example, the Carolina Capes (Ashton et al. 2001) (Fig. 2.2A). (On natural coastlines, shorelines from previous sea-level high stands, and other region-specific contingencies, will provide initial conditions that differ substantially from the straight-coastline initial condition in the model (Ashton and Murray 2006b). Such influences can act as finite-amplitude perturbations—starting points for the instability and self-organization that will define a significant initial wavelength, although the wavelength will tend to evolve subsequently.)

During this evolution, the shapes of the capes and bays remain nearly constant, in quasi-equilibrium with the forcing. The shape self-organizes into a configuration that minimizes gradients in alongshore sediment flux, given complex gradients in wave-shadowing effects—which themselves depend on the shapes. This configuration also prevents the long-term growth of new shoreline perturbations. In the bay between two capes, shadowing from the capes creates a locally low-angle-dominated wave climate. Shadowing also plays a role on the flanks of the capes, although the reorientation of the local shoreline, relative to the original global orientation, also alters local wave climates. Those cape-flank shorelines face toward some of the globally high-angle waves, which are therefore locally low-angle. The combination of shadowing and shoreline reorientation leads to locally diffusive wave climates everywhere except at the cape tips (Ashton and Murray 2006b). Ashton and Murray (2006b) have used these model wave-climate trends as a prediction, and the trends on the Carolina Capes are consistent with them.

Thus, this model of shoreline evolution behaves in an actively anti-fractal manner; a dominant wavelength grows in a way that prevents the introduction of any other length scales. On the other hand, the pattern is self-similar in another sense. Because the shadowing and shoreline-orientation effects are scale independent, the resulting shapes of the capes and bays remains constant as they grow—temporal self-similarity. Processes and resulting patterns that are scale-independent are one of the hallmarks of the turbulence paradigm.

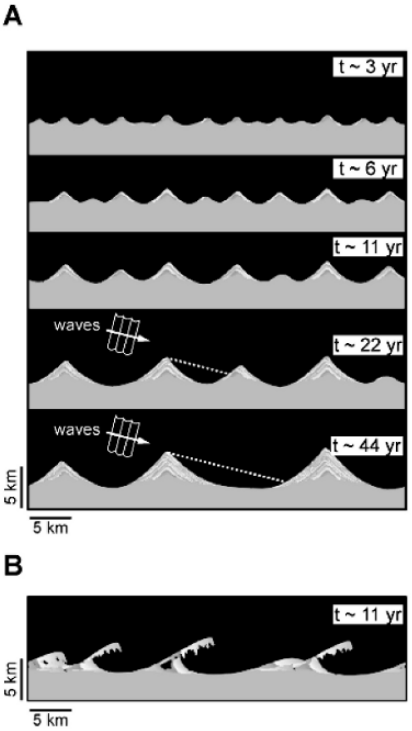


Fig. 2.6. Numerical model results, showing plan views. A: Shoreline evolution under a symmetric wave climate weighted toward high-angle waves, driven by wave shadowing. B: A snapshot under a strongly asymmetric wave climate (more waves approaching from the left). With both symmetric and asymmetric wave climates, the scale of the pattern continues to increase, leading to alongshore wavelengths on the order of 100 km after a timescale of millennia.

On natural capes other processes not included in the model (such as the coming and going of nearshore bars and rip channels) constantly introduce shoreline perturbations. Even if the diffusive local wave climates prevent the long-term growth of these perturbations, some days the waves approach from locally high angles. The repeated temporary growth and subsequent diffusion of perturbations will tend to create subtle shoreline variability on all scales (Pelletier and Turcotte 1999). Therefore, alongshore-transport processes could also explain the power-law scaling in shoreline changes observed on relatively small time scales along individual shoreline stretches (Tebbens and Nelson 2002). (Work is underway to test this hypothesis.)

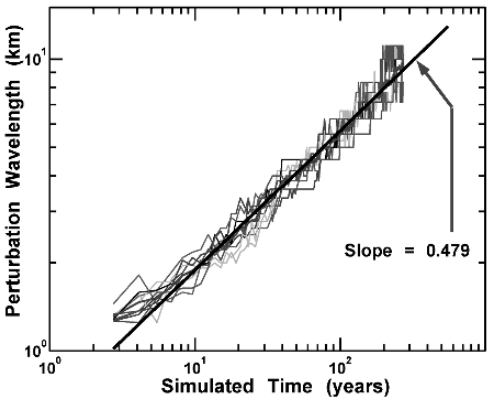


Fig. 2.7. Wavelength of capes in the model as a function of time, showing several runs (after Ashton and Murray 2006a).

The interactions in the coastline model are simple, even if the resulting behavior is complex. With the spatially extended domain there are potentially a large number of degrees of freedom—e.g. the position and orientation of each shoreline segment on the kilometer scale—even ignoring faster- and smaller-scale features such as nearshore bars. However, the capes that arise act as emergent structures that can be though of as constituting a smaller number of degrees of freedom: the cross-shore extent and alongshore position of each cape. The analysis of cape growth described above involves these emergent variables and the essentially deterministic interactions between them. (Even though the probabilistic treatment of deep-water wave-approach angles introduces stochasticity in the short-term forcing, in the longer term the time-averaged local wave climates are all that needs to be considered.) Thus, because of the emergence of large-scale variable and interactions, coastline evolution—including scaling aspects such as temporal self-similarity—need not be considered high dimensional or stochastic.

4 Discussion

The examples above show that large-scale landscape evolution involves various processes on many scales, and that landscape systems appear to exhibit different properties when viewed at different scales. At the scale of a grain in a river or a surf zone, or even of a gopher on a hillslope or a nearshore bar, the systems are very high dimensional (involving a large number of degrees of freedom) and stochastic, and simple holistic understanding doesn't seem to be within reach. However, basing models (conceptual, analytical and numerical) on interactions between larger-scale structures—such as those between channels and hillslopes or between different coastline features—is likely to lead to more-direct mechanistic explanations for landscape evolution. When described this way, spatially extended landscapes can still be high dimensional, although their behaviors can often be understood in terms of simple deterministic local interactions.

This emergent-phenomena approach, along with facilitating clearer explanations than do models based on the smallest-scale processes practical, may also produce more-accurate predictions (of typical system behaviors, not specific occurrences)—simulations that are more quantitatively reliable (Murray 2007; Murray et al. 2005; Werner 2003). Developing empirically based parameterizations of larger-scale relationships circumvents the risk that the interactions between various parameterizations of smaller-scale processes won't reproduce those larger-scale relationships realistically.

On the other hand, the power-law scaling that topography and river patterns exhibit indicates that scale-independent interactions play a large role in shaping the landscape. The two different approaches to studying landscapes sound incompatible; looking for scale-free interactions (the turbulence paradigm) seems inconsistent with treating interactions at different scales separately (emergent-phenomena paradigm). However, the examples examined here also show that the two approaches can be consistent and complementary; self-similar structures can be treated as emergent phenomena to be explained. Treating interactions at scales commensurate with elements of those structures—such as channels and hillslopes of various sizes—provides an effective way of studying the emergent scaling.

5 Conclusions

Even though spatially extended landscape systems can be viewed as high-dimensional, and they exhibit stochastic spatial and temporal patterns, essentially deterministic emergent interactions can often explain these dauntingly complex behaviors. Scaling analyses effectively characterize the stochastic self-similarity (affinity) of many of these patterns, and those scale-invariant properties can be effectively treated as emergent phenomena.

Acknowledgments

This paper reflects conversations with many people, chiefly Brad Werner, Chris Paola, and Peter Haff (who also provided a useful critique). The National Science Foundation (DEB 0507987) provided support.

References

Anderson, R. S. (2002) Modeling the tor-dotted crests, bedrock edges, and parabolic profiles of high alpine surfaces of the Wind River Range, Wyoming: Geomorph., 46, p. 35-58.

Ashton, A., List, J., Murray, A. B., and Ferris, A., 2003a, *Investigating Links Between Erosional Hot Spots and Alongshore Sediment Transport Using Field Measurements and Simulations*: Coastal Sediments '03, p. ISBN: 981-238-422-7.

Ashton, A., and Murray, A. B. (2006-a) Consequences of an instability in shoreline shape due to wave-driven alongshore sediment transport. Part 1: Theoretical and numerical investigations: J. Geophys. Res. 111, F04011, doi:10.1029/2005JF000422.

— (2006-b) Consequences of an instability in shoreline shape due to wave-driven alongshore sediment transport. Part 2: Wave climate analysis and comparisons to nature.: J. Geophys. Res. 111, F04011, doi:10.1029/2005JF000423.

Ashton, A., Murray, A. B., and Arnoult, O. (2001) Formation of coastline features by large-scale instabilities induced by high-angle waves: Nature, 414, p. 296-300.

Ashton, A., Murray, A. B., and Ruessink, G. B., 2003b, *Initial tests of a possible explanation for alongshore sandwaves on the Dutch Coast*: 3rd Internation Symposium on River, Coastal and Estuarine Morphodynamics, p. 320-330, ISBN: 90-805649-6-6.

Chase, C. G. (1992) Fluvial landsculpting and the fractal dimension of topography: Geomorphology, 5, p. 39-57.

Coulthard, T. J., Macklin, M. G., and Kirkby, M. J. (2002) A cellular model of holocene upland river basin and alluvial fan evolution: Earth Surf. Proc. Landforms, 27, p. 269-288.

Falques, A. (2003) On the diffusivity in coastline dynamics: Geophys. Res. Lett., 30, p. 2119, doi:10.1029/2003GL017760.

Furbish, D. J. (1993) Flow structure in a bouldery mountain stream with complex bed topography: Water Resour. Res., 29, p. 2249-2264

Gagnon, J. S., Lovejoy, S., and Schertzer, D. (2003) Multifractal surfaces and terrestrial topography: Europhys. Lett., 62, p. 801-807.

Gupta, V. K., and Waymire, C. E. (1989) Statistical self-similariy in river networks parameterized by elevation: Water Resour. Res., 25, p. 463-476.

Howard, A. D. (1994) A detachment-limited model of drainage basin evolution: Water Resour. Res., 39, p. 2261-2285.

Kessler, M., Murray, A. B., Werner, B. T., and Hallet, B. (2001) A model for sorted circles as self-organized patterns: J. Geophys. Res., 106, p. 13,287-13,306.

Kessler, M., and Werner, B. T. (2003) Self-organization of sorted patterned ground: Science, 299.

Komar, P. D. (1998) *Beach Processes and Sedimentation*: Upper Saddle River, New Jersey, Simon & Schuster, 544 p.

Kyungsoo, Y., Amundson, R., Heimsath, A. M., and Dietrich, W. E. (2005) Process-based model linking pocket gopher (Thomomys bottae) activity to sediment transport and soil thickness Geolgoy, 33, p. 917-920.

Leopold, L. B., and Wollman, M. G. (1957) River channel patterns, braided, meandering and straight, *Professional Paper*, US Geological Survey.

List, J. H., and Ferris, A. S., 1999, *Large-scale shoreline response to storms and fair weather*: Coastal Sediments '99, p. 1324-1338.

Lorentz, E. N. (1963) Deterministic, non-periodic flow: J. Atmos. Sci., 20, p. 130-141.

Mandelbrot, B. B. (1983) The fractal geometry of nature: Am. J. Phys. , 51.

Miall, A. D. (1992) Exxon global cycle chart: an event for every occasion?: Geology, 20, p. 787-790.

Murray, A. B. (2003) Contrasting the Goals, Strategies, and Predictions Associated With Simplified Numerical Models and Detailed Simulations, *in* Iverson, R. M., and Wilcock, P. R., eds., *Prediction in Geomorphology*, AGU Geophys. Monogr., p. 151-165.

— (2007) Reducing model complexity for explanation and prediction: Geomorph., in press.

Murray, A. B., and Ashton, A., 2004, *Extending a 1-line modeling approach to explore emergent coastline behaviors*: Coastal Engineering 2004, p. 2035-2047.

Murray, A. B., Coco, G., Green, M., Hume, T. M., and Thieler, E. R., 2005, *Different approaches to modeling inner shelf sorted bedforms*: 4th International Symposium on River, Coastal and Estuarine Morphodynamics, p. in press.

Murray, A. B., and Paola, C. (1994) A cellular model of braided rivers: Nature, 371, p. 54-57.

— (1997) Properties of a cellular braided stream model: Earth Surf. Proc. Landf., 22, p. 1001-1025.
— (2003) Modelling the effect of vegetation on channel pattern in braided rivers: Earth Surf. Proc. Landforms, 28, p. 131-143.
Murray, A. B., and Thieler, E. R. (2004) A new hypothesis for the formation of large-scale inner-shlef sediment sorting and 'Rippled Scour Depressions': Continental Shelf Res., 24, p. 295-315.
Parker, G. (1978) Self-formed straight rivers with equilibrium banks and mobile bed. Part 2. The gravel river: Jornal of Fluid Mechamics, 89, p. 127-146
Pelletier, J. D., and Turcotte, D. L. (1999) Self-Affine time series II: Applications and Models: Advances in Geophysics, 40, p. 91-166.
Reneau, S. L., and Dietrich, W. E. (1991) Erosion rates in the southern Oregon Coast Range: Evidence for an Equilibrium between Hillslope Erosion and Sediment Yield: Earth Surf. Proc. Landforms, 16, p. 307-322.
Ritter, D. F., Kochel, R. C., and Miller, J. R. (2001) *Process Geomorphology*, -Hill Science/Engineering/Math.
Rodriguez-Iturbe, I., Marani, M., Rigon, R., and Rinaldo, A. (1994) Self-organized river basin landscapes: Fractal and multifractal characteristics: Water Resour. Res., 30, p. 3531-3539.
Rogriguez-Iturbe, I., and Rinaldo, A. (1997) *Fractal River Basins*: Cambridge, Cambridge University Press, 564 p.
Rundle, J., Klein, W., and Turcotte, D. L. (1996) *Ruduction and predictability of natural hazards*, Westview Press.
Sapozhnikov, V. B., and Foufoula-Georgiou, E. (1996) Self-afinity in braided rivers: Water Resources Research., 32, p. 1429-1440
— (1999) Horizontal and veritcal self-organization of braided rivers toward a critical state: Water Resources Research, 35, p. 843-852.
Sapozhnikov, V. B., Murray, A. B., Paola, C., and Foufoula-Georgiou, E. (1998) Validation of braided-stream models: spatial state-space plots, self-affine scaling, and island shapes: Water Resour. Res, 34, p. 2353-2364.
Schmeeckle, M. W., and Nelson, J. M. (2003) Direct numerical simulation of bedload transport using a local, dynamic boundary condition: Sedimentology, 50.
Schmeeckle, M. W., Nelson, J. M., Pitlick, J., and Bennet, J. P. (2001) Interparticle collision of natural sediment grains in water: Water Resour. Res., 37, p. 2377–2392.
Sklar, L. S., and Dietrich, W. E. (2001) Sediment and rock strength controls on river incision into bedrock Geology, 29, p. 1087-1090.
Stark, C. P. (2006) A self-regulating model of bedrock river channel geometry: Geophy. Res. Lett., 33, p. L04402.
Strudley, M. W., Murray, A. B., and Haff, P. K. (2006) Emergence of pediments, tors and piedmont junctions from a bedrock weathering-regolith thickness feedback: Geology, 43, p. 805-808.
Tebbens, S. F., and Nelson, E. (2002) Wavelet analysis of shoreline change on the Outer Banks of North Carolina: Proc. of the Nat. Acad. of Sci., 99 (Suppl. 1), p. 2554-2560.
Thomas, R., and Nicholas, A. P. (2002) Simulation of braided river flow using a new cellular routing scheme: Geomorph., 43, p. 179-195.
Turcotte, D. L. (1997) *Fractals and Chaos in Geology and Geophysics*: Cambridge, Cambridge University Press, 412 p.
Turcotte, D. L., Malamud, B. D., Guzzetti, F., and Reichenbach , P. (2002) Self-organization, the cascade model, and natural hazards: Proceedings of the National Academy of Science, 99, p. 2530-2537.

Veitzer, S., and Gupta, V. K. (2000) Random self-similar river networks and derivations of generalized Hortons laws in terms of statistical simple scaling: Water Resour. Res., 36, p. 1033-48.

Werner, B. T. (1999) Complexity in natural landform patterns: Science, 284, p. 102-104.

— (2003) Modeling landforms as self-organized, hierarchical dynamic systems, *in* Iverson, R. M., and Wilcock, P., eds., *Prediction in Geomorphology*, AGU Geophys. Monogr., p. 131-150.

Werner, B. T., and Fink, T. M. (1993) Beach cusps as self-organized patterns: Science, 260, p. 968-971.

Werner, B. T., and Hallet, B. (1993) Numerical simulation of self-organized stripes: Nature, 361, p. 142-145.

Whipple, K. X., Hancock, G. S., and Anderson, R. S. (2000) River incision into bedrock: Mechanics and relative efficacy of plucking, abrasion, and cavitation: GSA Bull., 112, p. 490-503.

Whipple, K. X., and Tucker, G. E. (1999) Dynamics of the stream-power river incision model: Implications for height limits of mountain ranges, landscape response timescales, and research needs: Jour. Geophys. Res., 104, p. 17661-17674

Willgoose, G., Bras, R. L., and Rodriguez-Iturbe, I. (1991) A coupled channel network growth and hillslope evolution model. I. Theory: Water Resour. Res., 27, p. 1671-1684.

Yoo, K., Amundson, R., Heimsath, A. M., and Dietrich, W. E. (2005) Process-based model linking pocket gopher (*Thomomys bottae*) activity to sediment transport and soil thickness: Geology, 33, p. 917-920.

3 Effects of Systematic and Random Errors on the Spatial Scaling Properties in Radar-Estimated Rainfall

Gabriele Villarini[1], Grzegorz J. Ciach[2], Witold F. Krajewski[3], Keith M. Nordstrom[4], and Vijay K. Gupta[5]

[1] University of Iowa, IIHR-Hydroscience & Engineering, gabriele-villarini@uiowa.edu
[2] University of Iowa, IIHR-Hydroscience & Engineering, g-ciach@uiowa.edu
[3] University of Iowa, IIHR-Hydroscience & Engineering, witold-krajewski@uiowa.edu
[4] Northwest Research Associates, nordstro@cires.colorado.edu
[5] University of Colorado, guptav@cires.colorado.edu

Abstract. Spatial scaling properties of precipitation fields are often investigated based on radar data. However, little is known about the effects of the considerable uncertainties present in radar-rainfall products on the estimated multifractal parameters. The basic systematic factors that can affect the results of such analyses include the selection of a Z-R relationship, the rain/no-rain reflectivity threshold, and the distance from the radar. In addition, the inevitable random errors can strongly distort the radar-based scaling characteristics. The authors examine these problems using high-resolution radar-rainfall maps generated with different parameters based on the Level II data from a WSR-88D radar in Kansas, USA. To investigate the effects of random errors, the authors convolute the radar-rainfall with a multiplicative random uncertainty factor, and discuss the effects of different magnitudes and spatial dependences in such an uncertainty process. This study shows that the sensitivity of the multifractal analyses on some of the considered factors is strong and can even dominate the results.

1 Introduction

Knowledge of the precipitation structure at small spatial scales is important for realistic modeling of hydrometeorological phenomena at the larger scales. A good example here is the interaction between land and atmosphere hydrologic processes. The surface and subsurface processes are driven by the rainfall changes at much smaller distances than the scales of the corresponding weather patterns. The increasing realism of the water cycle models used in hydroclimatological studies requires information on the rainfall input at increasingly higher spatial resolutions. Such precipitation fields cannot be obtained with the General Circulation Models (GCMs) or even the Local Area Models (LAMs). Similarly, the global precipitation products obtained from satellites do not have the required resolution and reliability. This situation creates strong demand for the development of effective downscaling methods, i.e. the empirically substantiated techniques to distribute the coarse resolution area averaged amounts of rainfall among the areal elements of smaller scales. Weather radars, despite their limitations, provide invaluable data to develop such methods due to

their relatively high spatial sampling resolution and large coverage area of the observed precipitation systems. Although direct measurements from rain gauges and disdrometers provide more accurate data on the point rainfall, the radar is the only instrument capable of bridging the gap between the weather pattern scales (about 1000s km^2 and above) and the scales considered in modern hydrology (about 1 km^2 and below).

A promising approach to characterize precipitation processes across a wide range of space and time scales is the concept of scale invariance. Based on such methods, useful statistical information about the process at the scales much smaller than the original data resolution can be derived, provided that relatively simple power-law relationships are fulfilled across an adequate range of scales (e.g., Lovejoy and Mandelbrot 1985; Waymire 1985; Gupta and Waymire 1990; Gupta and Waymire 1993; Tessier, Lovejoy, and Schertzer 1993; Over and Gupta 1994; Lovejoy and Schertzer 1995; Perica and Foufoula-Georgiou 1996). High resolution rainfall properties obtained from various downscaling models based on the scale invariance assumptions can be used, for instance, as input to distributed hydrological models in order to improve the predictions of water and energy exchange between the Earth surface and the atmosphere, both at the local and global scales. As a rule, spatial scaling properties of rainfall have been studied based on weather radar data because they have two unique advantages in comparison with all other precipitation observations: (1) continuous coverage of large observation areas and (2) relatively high spatial resolution. However, so far, little attention has been given to the effects of the uncertainties in the estimates of radar-rainfall (RR) on the computed scaling characteristics. How the RR errors propagate into the obtained scaling properties and how they affect the parameters in the multifractal models of spatial precipitation structure, still remain open questions (e.g., Krajewski, Anagnostou, and Ciach 1996; Harris, Seed, Menabde, and Austin 1997). We address these questions in this analysis.

The factors responsible for the large discrepancies between RR and the ground reference (GR) measurements have been discussed in numerous studies. The GR used in such comparisons is usually based on rain gauges that deliver approximations of the true rainfall corresponding to the evaluated RR estimates. See Krajewski and Smith (2002) for a recent discussion of issues involved in reliable verification of RR products, and Krajewski and Ciach (2003) for a comprehensive literature review on their uncertainties. Most of radar-rainfall uncertainty studies are either qualitative or focused on a single error source in isolation of other factors. Therefore, much of our efforts have been devoted to the development of a comprehensive quantitative RR uncertainty model that could remedy this situation (e.g., Ciach and Krajewski 1999a; Ciach and Krajewski 1999b; Ciach, Morrissey, and Krajewski 2000). Although this work is still far from being completed, it has already provided us with a considerable amount of new insight into the RR error structure. This allows us analyzing the possible consequences of RR errors in different applications. In this study, we discuss the sensitivity of multifractal analyses of RR fields with respect to the inherent RR uncertainties.

Our study builds on earlier efforts by Krajewski et al. (1996) and Harris et al. (1997). In the former study, we discussed the effects of the uncertainties on spatial

rainfall characteristics using two different RR estimation algorithms and an idealistic simulation of the radar measurement process. In the latter study, the authors investigate the uncertainties in the estimated multifractal parameters for rainfall time series based on collocated data from rain gauges and a vertically-pointing radar. However, as recently noted by Nykanen and Harris (2003), "even radar data of the best quality is subject to artifacts […] further study is needed to better quantify the effects on multiscaling parameter estimates."

In this paper, we use the Level II reflectivity data from a Weather Radar Surveillance Radar – 1988 Doppler (WSR-88D) station in Kansas, USA, to investigate the effects of systematic and random distortions in RR on the estimation of spatial multifractal characteristics. The systematic factors considered here are the selection of a Z-R relationship, the rain/no-rain reflectivity threshold, and the distance from the radar. To investigate the effects of random errors, we convolute the estimated RR fields with a multiplicative random uncertainty factor that has characteristics similar to the real random error process in RR. We consider the effects of different magnitudes and spatial dependences in the error process on the spatial multifractal estimates. Such sensitivity analysis helps us to identify the most important problems and set priorities for future investigation.

This work is motivated by the need to make rainfall modelers, and especially users of rainfall downscaling models, aware of the issue. Biased and unrealistic downscaling relationships may distort results of subsequent hydrologic prediction and modeling, and the conclusions of such studies can be highly misleading. RR products are also applied to develop and validate innovative theories of the dynamic behavior of rainfall (e.g., Nordstrom and Gupta 2003). The effects of the systematic and random uncertainties in RR estimates should be thoroughly recognized and accounted for in all such analyses.

This study is structured as follows. In the next section we present a brief outline of the moment scaling theory used here. In Section 3 we provide a description of the data sample. In Section 4 we illustrate the effects of systematic and random errors on the multifractal scaling functions. Section 5 summarizes the important points made in this article and concludes the paper.

2 Multifractal Moment Scaling

Different concepts of scale invariance in geophysical phenomena, including the spatiotemporal rainfall process, have been extensively studied for more than two decades already. In particular, the occurrence of multifractal behavior in rainfall has been reported in many papers (e.g., Lovejoy and Schertzer 1995; and references therein). By definition, a spatiotemporal random process X has multifractal properties if its probability distribution across a range of averaging scales changes according to specific power-law functions (Lovejoy and Schertzer 1995). This property is usually investigated in its mathematically equivalent form of multiple moment-scaling law (Lovejoy and Schertzer 1995) stating that, for the qth moment order:

$$E\left\{X^q\left(\lambda L_0\right)\right\} = \lambda^{-K(q)} E\left\{X^q\left(L_0\right)\right\} \tag{1}$$

where $E\{\cdot\}$ is the operator of statistical expectation, L_0 is the reference scale (usually the finest resolution scale), λ is the spatial scaling factor, and $K(q)$ is the scaling function.

To estimate the scaling function $K(q)$ for a given data sample, it is convenient to take the logarithm of both sides in Eq. 1. The multifractal scaling behavior is manifested by $\log(E\{X^q(\lambda L_0)\})$ being a linear function of $\log(\lambda)$ for each fixed moment order. Then, the slope of this linear function, for a given q, is the value of the scaling function $K(q)$. For the first moment, the scaling function $K(1)=0$ because the unconditional spatial average of the rainfall itself does not depend on the averaging scale. According to previous studies, the scaling function, $K(q)$, for different moment orders, q, is influenced by different levels of rainfall. For larger moment orders, the slopes of $\log(E\{X^q(\lambda L_0)\})$ versus $\log(\lambda)$ lines are mostly determined by the higher rainfall values and can be dominated by few extreme observations (Kumar, Guttarp, and Foufoula-Georgiou 1994). On the other hand, for moment orders below the value of $q=1$, the slopes are mostly affected by measurement noise, which can obstruct distinguishing small values from zeros (Harris, Menabde, Seed, and Austin 1996; Harris et al. 1997).

The final results of the moment scaling analysis outlined above are empirical estimates of the scaling function, $K(q)$, obtained in several different scenarios. Apart from considering various conditions of the RR estimation scheme and the random uncertainty structure, we perform our analysis for two distinctly different types of precipitation systems: convective and stratiform. The former tends to be organized at the scale of individual storm cells, and it is characterized by strong gradients, where peaks are surrounded by lower values of precipitation. The latter tends to be more organized on a larger scale and is in general smoother. We show that these differences in spatial variability are strongly reflected in the shapes of the scaling functions.

3 Data Sample and Preliminary Analysis

Our analysis is based on data obtained from the WSR-88D radar station in Wichita (KICT), Kansas, USA. In this study, we considered two long-lasting rainfall events that have distinctly different spatial characteristics. The first, called from now on Storm 1, occurred on 25 April 1999 between 00:06 and 20:01 UTC and it is mostly stratiform. The second event, called from now on Storm 2, is mostly convective, characterized by squall-lines crossing the radar domain, and it started on 27 September 1999 at 19:03 UTC and ended on 28 September 1999 at 17:05 UTC. In Fig. 3.1, we show examples of radar scans from these two situations, the stratiform in the upper panel and the convective in the lower panel, together with the areal domains considered in this study. The analysis domain has the side length of 64 km. We exclude the center of the radar coverage to avoid echo contamination by ground clutter that could add additional uncertainty to our analysis. The finest scale of our analysis is 1 km by 1 km and roughly corresponds to the resolution of the radar data at the distances considered. The largest averaging scale in this analysis is 32 km by 32 km, therefore the value of the spatial scaling factor in Eq. 1 ranges from 1 to 32.

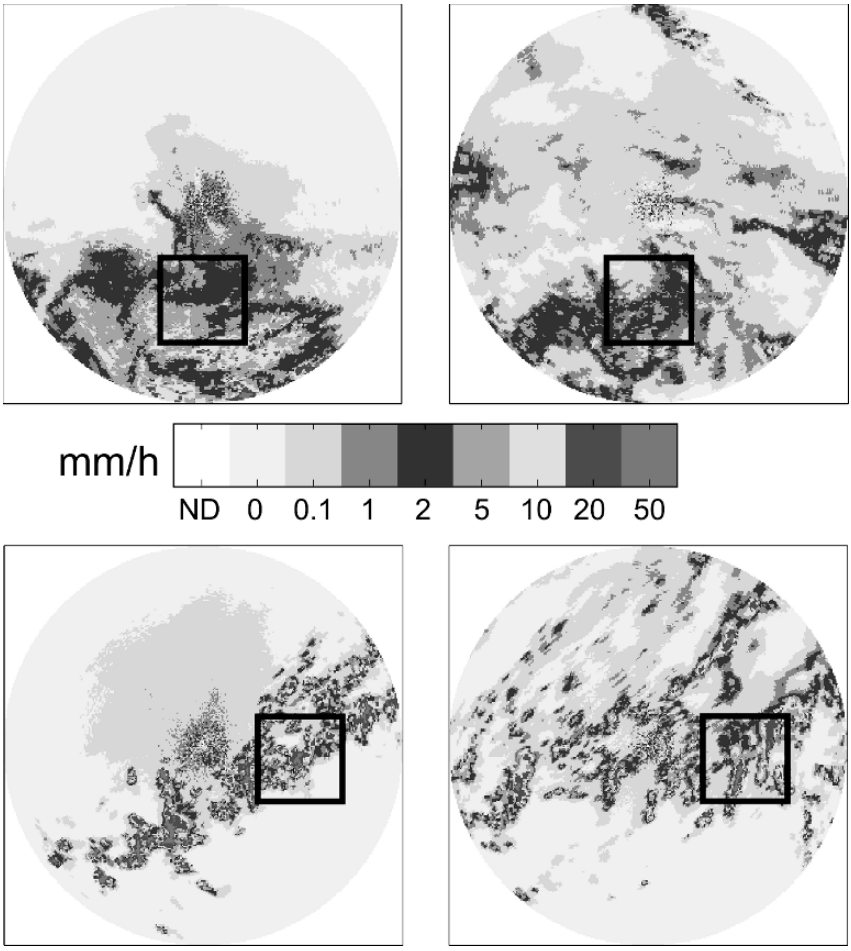


Fig. 3.1. Radar scans from the stratiform (upper panel) and convective (lower panel) events. The black square represents the domain considered in this study. (A color version of this figure appears between pages 196 and 197).

Rainfall rate (R, in mm/h) is obtained from the measured radar reflectivity (Z, in mm^6/m^3) using a power-law relationship, $Z = aR^b$, where a and b are its two adjustable parameters. In the WSR-88D stations in the United States, these two parameters are usually set to their standard values of $a=300$ and $b=1.4$ (Fulton, Breidenbach, Seo, and Miller 1998), and these specific values are used throughout this study, unless specified differently. After checking that neither storm is affected by the presence of anomalous propagation echoes, we converted the base scan reflectivities into RR intensities and transformed the data from polar to Cartesian coordinates using the nearest neighbor scheme. These data processing steps constitute a simple RR estimation algorithm. The obtained rain fields have a spatial resolution of 1 km by 1 km and the temporal resolution of 5–6 minutes.

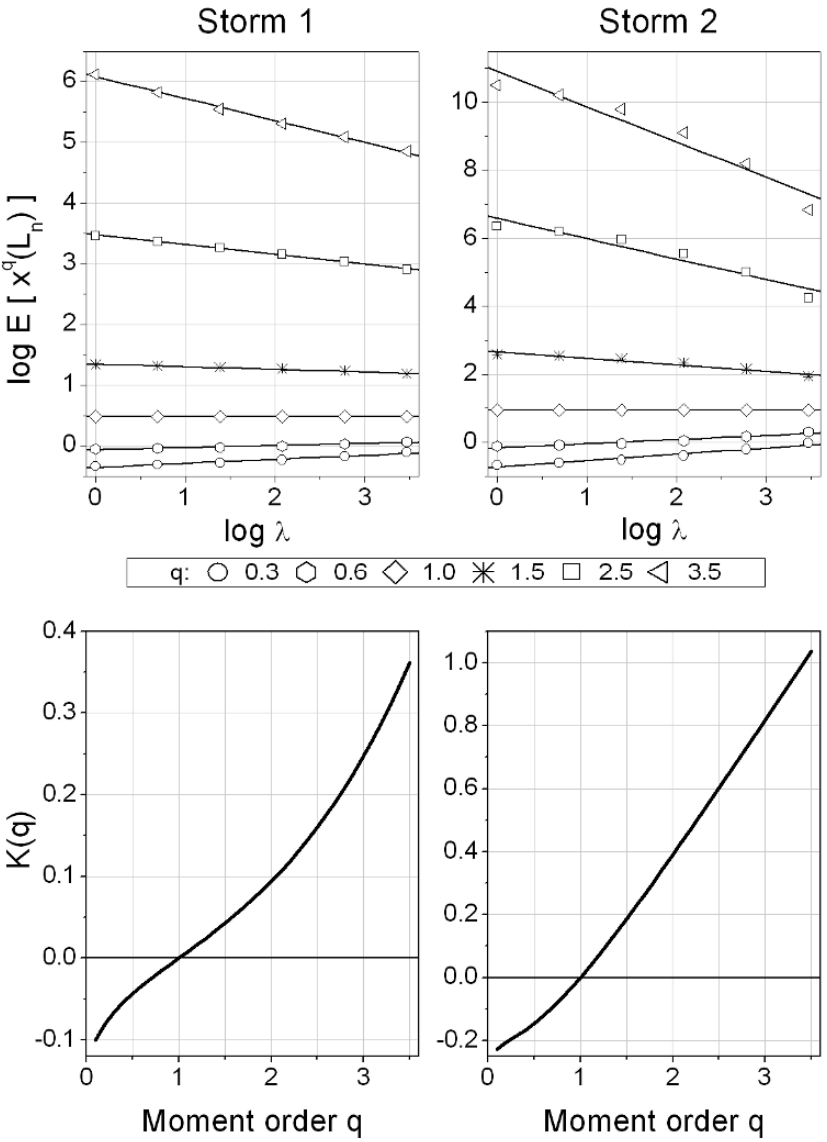


Fig. 3.2. Moment scaling plots (upper panels) and the corresponding scaling functions (lower panel) for Storm 1 and Storm 2.

To check the validity of the moment scaling hypothesis, we plotted the estimates $\log(E\{X^q(\lambda L_0)\})$ versus $\log(\lambda)$. As shown in Fig. 3.2 (upper panels), the points for each fixed value of the moment, q, are arranged fairly accurately along straight lines.

We can also notice a substantial difference in the slopes of the regression lines between the Storm 1 and Storm 2. To illustrate the analysis method that we use further on, we summarize this difference with the plots in the bottom panels in Fig. 3.2 presenting the corresponding estimates of the scaling functions, *K(q)*, for the two events.

4 Sensitivity Analysis

The radar reflectivity measurements are only statistically related to the RR intensities. This fundamental fact, as well as other factors, leads to inevitable discrepancies between RR and the corresponding true rainfall (e.g., Ciach and Krajewski 1999b; Krajewski and Smith 2002). However, little is known about the effects of these errors on the estimated spatial scaling properties of rainfall. Our goal is to show that the inherent uncertainties in RR can have a major impact on the estimated scaling characteristics and might even dominate the results. Below, we investigate the sensitivity of the scaling properties of RR fields to the systematic and random uncertainty factors.

4.1 Effect of systematic errors

In this section, we present the effects of different Z-R relationships, rain/no-rain thresholds and the distance from the radar.

First, we focus on the selection of the Z-R exponent, *b*. The value of the multiplicative parameter, *a*, is assumed to be constant and set to *300* because it has no influence on the scaling functions (Harris et al. 1997). In this study, we estimate the scaling functions, *K(q)*, using four different values of *1.2*, *1.4*, *1.6* and *1.8* of the *b* parameter. Fig. 3.3 (upper panels) shows that the selection of the exponent has a strong impact on the scaling functions, and that increasing of *b* always decreases the *K(q)* values. The changes of *b* affect the nonlinear conversion from *Z* to *R* and are more influential for higher order moments. This is reflected in the behavior of the corresponding scaling functions in Fig. 3.3, where the *K(q)* curves for different *b* values are more far apart for larger values of *q*. On the other hand, the scaling functions for the lower moment orders are much less sensitive to the selection of the Z-R exponent.

The next systematic effect investigated in this paper is the selection of the rain/no-rain threshold that represents the reflectivity value below which the signal is classified as no rain. We considered the threshold values of 10 and 20 dBZ, with the Z-R exponent set to its standard value of *b=1.4*. These results are presented in Fig. 3.3 in the middle panels. It is clear that, for both the stratiform and convective case, the threshold selection has an effect on the *K(q)* behavior at the moment orders smaller than 1 only, whereas the higher moment orders are not much affected by it. The reflectivity threshold changes the area of the RR field where it is equal to zero. Thus, the behavior of the scaling function at the *q* values smaller than 1 carries information about the zero-intermittency in the rainfall fields.

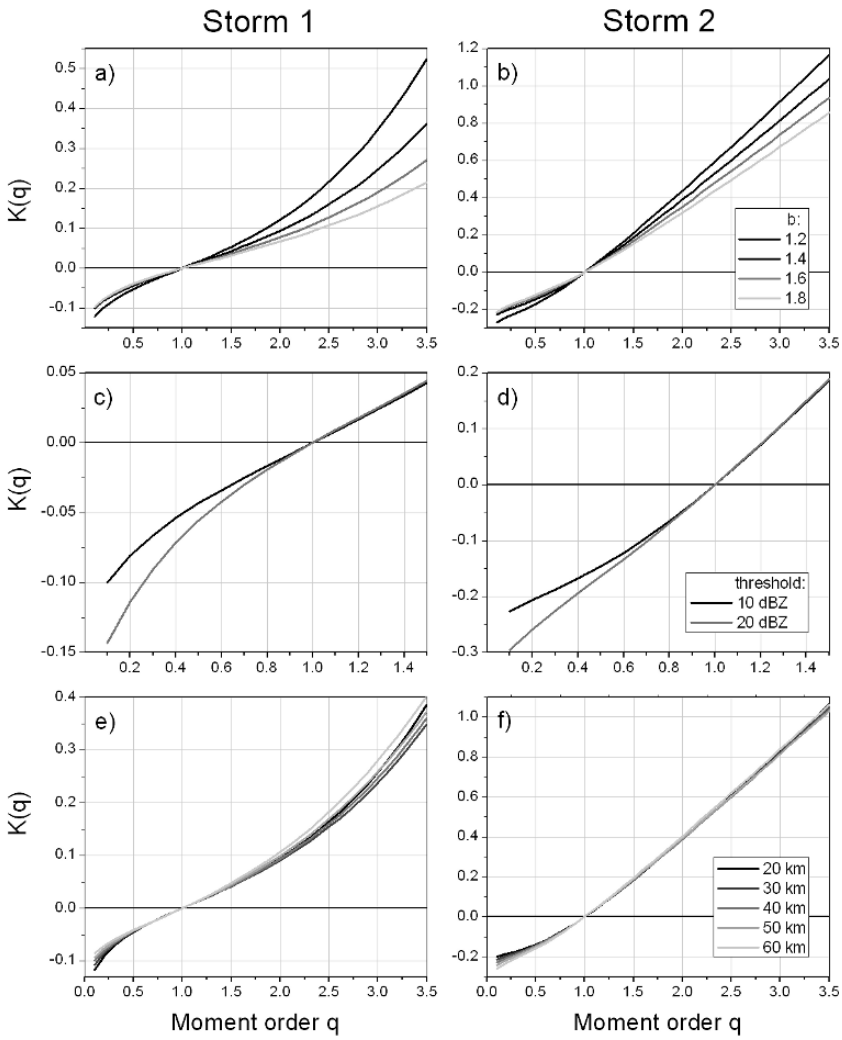


Fig. 3.3. Scaling functions for different values of the Z-R exponent *b* (upper panel), for different rain/no rain thresholds (middle panels) and for different distances from the radar site (bottom panels).

The last systematic factor that we examine in this study is the effect of the distance from the radar on the scaling functions. Range dependent biases have a strong influence on RR estimates and their hydrological applications (e.g., Smith, Seo, Baeck, and Hudlow 1996). The reflectivity threshold is set to 10 dBZ and the Z-R exponent to its standard value. In Fig. 3.3 in bottom panels, we present the estimates of the *K(q)* functions for different distances of the analysis domain. The values

reported in the legends represent the distance from the radar of the domain side that is closer to the radar location. This distance ranges from 20 km up to 60 km. For distances smaller than 20 km there is a significant amount of ground clutter in the radar echoes. On the other hand, for distances larger than 60 km, the farther side of the domain is at the distance above 120 km and the resolution of the basic radar data in the azimuth is much larger than 1 km due to the radar beam broadening. Therefore the range of the distances considered here is so limited. For this limited range (between about 50 and 90 km of the domain center), the distance has a relatively small impact on the scaling function for the moment orders larger than 1, especially for the convective case. For the stratiform case, the effect seems to be significant, but the dependence of the $K(q)$ changes on the distance is irregular. For the moment orders much smaller than 1, the effect seems to be both significant and of a regular character. This is probably associated with the systematic dependence of the zero-intermittency structure in RR fields on the distance from the radar. Interestingly, this dependence has an opposite direction for the stratiform and convective cases, however, we did not find any clear interpretation of this behavior.

4.2 Effect of random errors

The random uncertainties associated with RR estimates are often considered to be multiplicative and lognormally distributed (e.g., Smith and Krajewski 1993; Ciach and Krajewski 1999b; Ciach et al. 2000). The spatial dependence structure of these RR uncertainties is certainly an important factor responsible for the possible large differences between the scaling properties of RR and the corresponding true rainfall fields. However, estimating these dependences in the error process is a challenging problem and it still remains unexplored. For this reason, similarly to Nijssen and Lettenmaier (2004), we investigate in this study the possible significance of these dependencies by considering two scenarios: spatially correlated and uncorrelated random uncertainty fields imposed on the original RR fields.

The correlated random perturbation fields are generated starting from the typical spatially correlated standard Gaussian fields (mean $\mu=0$, standard deviation $\sigma=1$, and known spatial correlation function). These generic fields are transformed to the positively defined spatially correlated lognormal fields through an exponential function, $Y=c_1 \exp(c_2 X)$, where the parameters c_1 and c_2 allow controlling the expectation (always set to the value of 1) and variance of the outcome. Since the functional relationship between the correlation coefficients of the Gaussian and the corresponding lognormal variables is well known (e.g., Habib, Krajewski, and Ciach 2001), this way we can generate fields of lognormally distributed random variables with known spatial correlations. In this study, we discuss the case of strongly correlated error fields with the correlation distance of about 30 km. Multiplying the original RR fields estimated from the radar data by these perturbation fields (P) we obtain Monte-Carlo ensembles of the possible realizations of the uncertain rainfall estimates (RR_p). The ensemble sizes for this analysis are as big as 1000 perturbation fields for each of the considered cases of different random uncertainty characteristics. Note

that the perturbation scheme leaves the areas with zero rainfall (the zero-intermittence structure) unmodified.

The perturbation fields that we use here are two-dimensional stochastic processes with simple spatial dependency structure. Applying the procedure described in Section 2, we verified that, indeed, they do not exhibit any scaling properties. However, the perturbed RR fields still have the moment scaling properties defined and illustrated in Section 2, although the specific values in their scaling functions can differ from the original ones. This shows that the existence of spatial scaling properties themselves is robustly persistent in respect to non-scaling stochastic perturbations. However, their detailed quantitative characteristics can be sensitive to the random errors (just as they were shown to be sensitive to some of the systematic factors). Below, we investigate this sensitivity in more detail.

For each of the two cases of spatially uncorrelated and correlated lognormal perturbations, we considered three values their coefficient of variation, CV_e, equal to *0.05*, *0.5* and *1.0*. Fig. 3.4 shows the effect of these random uncertainties on the estimated scaling function in Storm 1, whereas Fig. 3.5 shows the same results for Storm 2. The panels on the left side of these two figures present the results for the uncorrelated error fields, and the panels on the right apply to the correlated error fields. The solid lines in each panel represent the *K(q)* functions for the original unperturbed RR fields. The results of the Monte-Carlo perturbation analysis are represented by the box-plots in the panels. The line in the middle of each box is the median of the 1000 realizations, the lower and upper edges of each box are the 25^{th} and 75^{th} quantiles, while the limits of the whiskers are the 5^{th} and 95^{th} quantiles.

The first pronounced feature visible in Figs 3.4, 3.5 is the existence of a systematic positive bias in the estimated scaling functions of the perturbed RR fields. As can be expected, the bias increases as the magnitude of the random error increases. It is practically negligible for $CV_e=0.05$ (*5%* of the relative standard error in RR), however, for the more realistic higher RR error levels, it becomes large. In addition, the bias grows with the increasing moment order. For example, at $q=3.0$ and for uncorrelated errors with $CV_e=0.5$, the scaling function is overestimated about two times. Thus, the non-scaling stochastic distortions imposed on a spatially scaling RR field do not obscure the existence of the scaling itself, but they can drastically change the quantitative characteristics of this scaling. Note also the rapidly broadening distributions of the sampling errors of the estimated scaling functions, *K(q)*, with the increasing values of the moment order and the random uncertainty level.

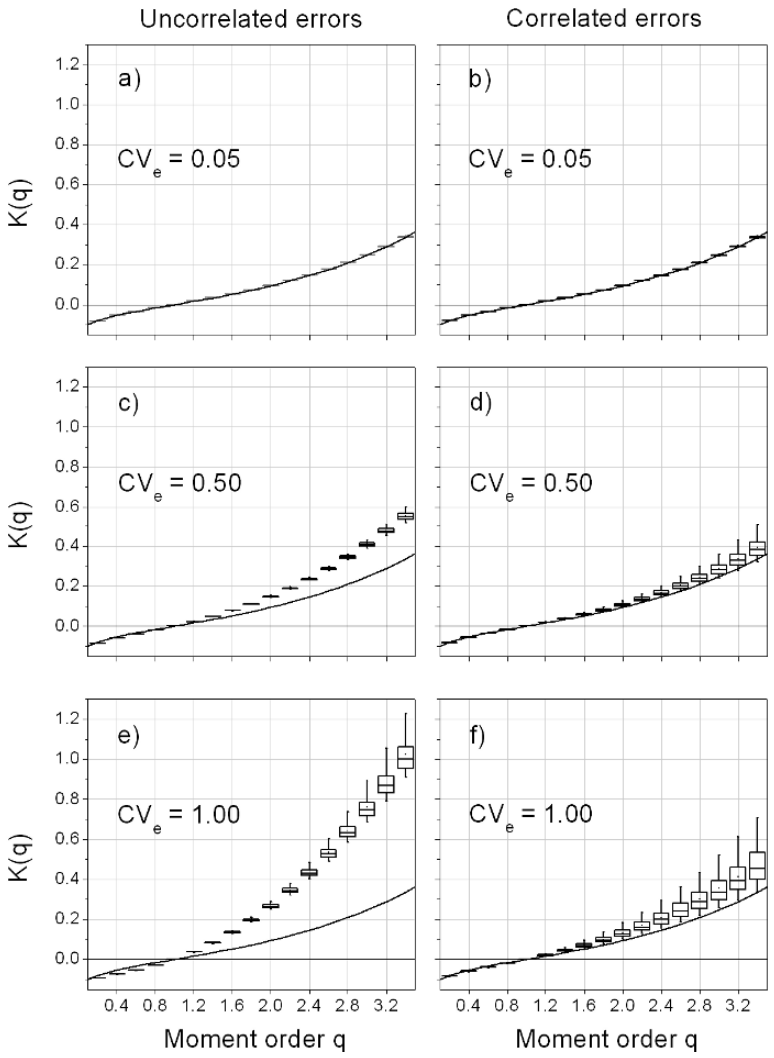


Fig. 3.4. Effects of random errors on the scaling properties in Storm 1.

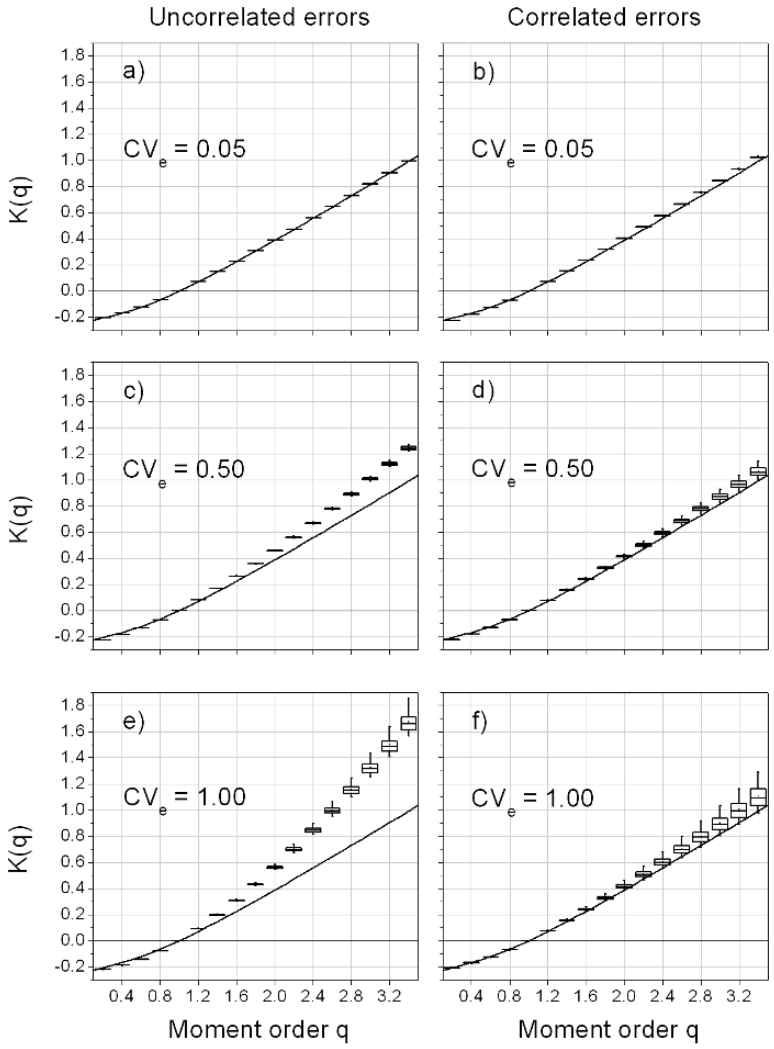


Fig. 3.5. Effects of random errors on the scaling properties in Storm 2.

Another interesting aspect of the results in Figs 3.4, 3.5 is the effect of the spatial correlations in the random error fields on the estimated scaling functions. For very low values of CV_e=*0.05*, this effect is not visible because, at such low error level, the effects of the uncertainties are negligible in any case. However, for the larger values of CV_e, the existence of spatial correlations in the random perturbations reduces the distortions in the scaling functions for both the stratiform and convective rain regimes. For the correlated errors (panels on the right side of Figs. 3.4–3.5) the

systematic departures of *K(q)* from the original RR are still evident, but their magnitudes are about five times smaller than for the uncorrelated errors. This shows that possible spatial dependences in the RR uncertainties can have considerable effects on the estimated scaling characteristics of the rainfall fields. Unfortunately, practically no information on these dependences is currently available.

The results in this section show that combinations of scaling and non-scaling fields can still exhibit scaling properties. Since the measurement errors often tend to have strong scaling features (Ciach 2003), it is even conceivable that the published spatial scaling characteristics based on RR fields are dominated by the scaling properties of the RR uncertainties rather than the scaling properties of the underlying true rainfall fields. If this is true, then any meaningful application of these characteristics might be questionable. Therefore, it is critical that we can devote more effort to investigate the spatial dependences in the RR error process.

Finally, comparing the corresponding panels between Figs. 3.4–3.5, one can evaluate the effects of random errors in the two distinctly different rainfall regimes. Generally, the observations discussed above are true for both Storm 1 and Storm 2. The only visible difference is that the effects of random errors on the scaling functions are relatively stronger in the stratiform regime than in the convective case.

5 Conclusions

For rational applications of radar-rainfall (RR) estimates (for example, the development of rainfall downscaling schemes), it is necessary to improve our understanding of the effects that the inevitable uncertainties in RR can have on the application outcomes. In this study, we demonstrated that the spatial multifractal scaling characteristics based on RR maps can be strongly distorted by the systematic and random errors in RR.

This analysis shows that various systematic factor have highly different influence on the estimated scaling functions. The rain/no-rain threshold level has a relatively insignificant effect, limited to the lower order moments. The distance from the radar site only weakly affects the scaling functions, at least for the ranges that we could examine here. The distance effect might be stronger over the larger distance range. However, the selection of the Z-R relationship exponent has a strong effect on the scaling characteristics, especially for the stratiform rainfall regime. Therefore, to study systematically the scaling properties of RR, it is important to know the Z-R exponent used for the conversion of radar reflectivity to rainfall intensity.

The effects of the random errors in RR depend strongly on the error magnitude and the existence of spatial correlations in the random error process. In addition, both the distortion level and the sampling uncertainty in the estimated scaling functions increase with the increasing moment orders. As expected, the distortions grow dramatically with increasing magnitude of the RR random error. On the other hand, the existence of spatial correlations in the random error field reduces considerably the distortions in the scaling functions of RR fields. Interestingly, our simulation shows that the scale invariance of the RR fields is a robust property that persists under

non-scaling multiplicative random perturbations (although the values of the estimated scaling function are sensitive to such random errors). This persistence, however, poses a fundamental question whether the observed spatial scaling in RR fields is not dominated by the possible scaling properties of the random error process that inevitably contaminates the RR estimates.

There are several directions that we would like to address in our future research. First of all, there is a need to develop a more realistic and empirically supported RR uncertainty model including the possible spatiotemporal scaling properties of the error process. Next, implementation of more robust scaling description methods is advisable in order to replace the moment scaling analysis, which, although the most popular in the published literature, may sometime yield spurious results (Veneziano, Moglen, and Bras 1995). Also, the differences between the stratiform and convective cases shown in this study indicate that the extensive investigation of large data samples containing a variety of rainfall regimes could provide more insight into the dependences between spatial scaling properties of RR fields and the climatological characteristics of the observed precipitation systems. Finally, a question should be posed about the values of the Z-R exponents in different synoptic conditions that could result in RR scaling functions to be close to the unknown scaling functions of the underlying true rainfall fields. Direct examination of this question requires accurate measurements of the true rainfall over a broad range of spatial resolutions and is beyond the capabilities of the currently existing rain gauge networks.

6 Acknowledgements

The authors acknowledge the support of the National Science Foundation through Grants EAR-0309644 and EAR-0450320. The first author was supported by NASA Headquarters under the Earth Science Fellowship Grant NGT5 - NNX06AF23H.

References

Ciach, G.J. and Krajewski, W.F. (1999a) On the estimation of radar rainfall error variance. Adv. Water Resour. 22(6), 585-595.

Ciach, G.J. and Krajewski, W.F. (1999b) Radar-rain gauge comparisons under observational uncertainties. J. Appl. Meteorol. 38, 1519-1525.

Ciach, G.J., Morrissey, M.L. and Krajewski, W.F. (2000) Conditional bias in radar rainfall estimation. J. Appl. Meteorol. 39(11), 1941-46.

Ciach, G.J. (2003) Local random errors in tipping-bucket raingauge measurements. J. Atmos. Oceanic Technol. 20, 752-759.

Fulton, R.A., Breidenbach, J.P., Seo, D.-J. and Miller, D.A. (1998) WSR-88D rainfall algorithm. Weather and Forecasting. 13, 377-395.

Gupta, V.K. and Waymire, E. (1990) Multiscaling properties of spatial rainfall and river flow distributions. J. Geophys. Res. 95(D3), 1999-2009.

Gupta, V.K. and Waymire, E. (1993) A statistical analysis of mesoscale rainfall as a random cascade. J. Appl. Meteorol. 32, 251-267.

Habib, E., Krajewski, W.F. and Ciach, G.J. (2001) Estimation of rainfall inter-station correlation. J. Hydrometeor. 2, 621-629.

Harris, D., Menabde, M., Seed, A.W. and Austin, G. (1996) Multifractal characterization of rain fields with a strong orographic influence. J. Geophys. Res. 101(D21), 26405-26414.

Harris, D., Seed, A.W., Menabde, M. and Austin, G. (1997) Factors affecting multiscaling analysis of rainfall time series. Nonlinear Processes in Geophysics. 4, 137-155.

Krajewski, W.F., Anagnostou, E.N. and Ciach, G.J. (1996) Effects of the radar observation process on inferred rainfall statistic. J. Geophys. Res. 101(D21), 26493-26502.

Krajewski, W.F. and Smith, J.A. (2002) Radar hydrology: rainfall estimation. Adv. Water Resour. 25, 1387-1394.

Krajewski, W.F. and Ciach, G.J. (2003) Towards probabilistic quantitative precipitation WSR-88D algorithms: Preliminary studies and problem formulation, report, 59 pp., NWS Office of Hydrologic Development, Silver Spring, Maryland.

Kumar, P., Guttarp, P. and Foufoula-Georgiou, E. (1994) A probability-weighted moment test to assess simple scaling. Stochastic Hydrology and Hydraulics. 8, 173-183.

Lovejoy, S. and Mandelbrot, B. (1985) Fractal properties of rain, and a fractal model. Tellus Ser. A. 37, 209-232.

Lovejoy, S. and Schertzer, D. (1995) Multifractals and rain. In: Z. Kundzewicz (Ed*.), New uncertainty Concepts in Hydrology and Water Resources*. Cambridge University Press, Cambridge, UK, pp. 62-103.

Nijssen, B. and Lettenmaier, D.P. (2004) Effect of precipitation sampling error on simulated hydrological fluxes and states: anticipating the Global Precipitation Measurement satellites. J. Geophys. Res. 109, D02103, doi:10.1029/2003JD003497.

Nordstrom, K. and Gupta, V.K. (2003) Scaling statistics in a critical, nonlinear physical model of tropical oceanic rainfall. Nonlinear Processes in Geophysics. 10, 531-543.

Nykanen, D.K. and Harris, D. (2003) Orographic influences on the multiscale statistical properties of precipitation. J. Geophys. Res. 108(D8), 8381, doi:10.1029/2001JD001518.

Over, T.M. and Gupta, V.K. (1994) Statistical analysis of mesoscale rainfall: dependence of a random cascade generator on large-scale forcing. J. Appl. Meteorol. 33, 1526-1542.

Perica, S. and Foufoula-Georgiou, E. (1996) Linkage of scaling and thermodynamic parameters of rainfall: results from midlatitude mesoscale convective systems. J. Geophys. Res. 101(D3), 7431-7448.

Smith, J.A. and Krajewski, W.F. (1993) A modeling study of rainfall rate-reflectivity relationships. Water Resour. Res. 29(8), 2505-2514.

Smith, J.A., Seo, D.-J., Baeck, M.L. and Hudlow, M.D. (1996) An intercomparison study of NEXRAD precipitation estimates. Water Resour. Res. 32(7), 2035-2045.

Tessier, Y., Lovejoy, S. and Schertzer, D. (1993) Universal multifractals: theory and observations for rain and clouds. J. Appl. Meteorol. 32, 223-250.

Veneziano, D., Moglen, G.E. and Bras, R.L. (1995) Multifractal analysis: pitfalls of standard procedures and alternatives. Physical Review E. 52, 1387-1398.

Waymire, E.C. (1985) Scaling limits and self-similarity in precipitation fields. Water Resour. Res. 21(8), 1271-1281.

4 Nonlinear Dynamics in the Earth's Magnetosphere

D.N. Baker[1], A.J. Klimas[2], and D. Vassiliadis[3]

[1] Laboratory for Atmospheric and Space Physics, University of Colorado at Boulder, Daniel.baker@lasp.colorado.edu

[2] National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, Maryland, 20771

[3] ST at NASA/GSFC, Goddard Space Flight Center, Greenbelt, MD 20771

Abstract. Observational evidence and numerical modeling demonstrate that substorms are a global, coherent set of processes within the magnetosphere and ionosphere. This knowledge supports the view that magnetospheric substorms are a configurational instability of the coupled global system. The magnetosphere progresses through a specific sequence of energy-loading and stress-developing states until the entire system suddenly reconfigures. The energy loading-unloading sequence is the basis of nonlinear dynamics models that have been successful in describing the essential behavior of substorms without invoking detailed treatments of the internal substorm instability mechanism. Recent results in data analysis and modeling show that powerful concepts in statistical physics and applied mathematics can be incorporated into space plasma physical research. A number of methods for modeling complex systems, data assimilation, system estimation, and predictive methods have been applied to the recent wealth of ionospheric and magnetospheric data. Nonlinear prediction schemes, for example, have greatly improved space weather forecasting and in most instances they remain more accurate and faster than physics-based models. From the standpoint of basic physical understanding, self-organized criticality recently has been used to describe scale-free avalanching phenomena observed in space plasma domains such as the plasma sheet and its low-altitude extension in the ionosphere. Self-organized criticality arises in the bursty transport of magnetic energy (flux) from the tail lobes, through the plasma sheet, and out of the mid-tail region where reconnection occurs. Multi-point measurement techniques and missions are being developed to determine the spatial and temporal development of the global phenomena that constitute energy dissipation within the magnetosphere.

1 Introduction

Magnetospheric substorms represent a global interaction between the solar wind, the magnetosphere, and the ionosphere (see Fig. 4.1a). Energy extracted from the solar wind is episodically stored in the magnetosphere, with a large fraction of the energy often being in the form of excess magnetic flux in the magnetotail lobes (Fig. 4.1b). This stored energy is then explosively dissipated in the near-Earth, nightside region at substorm onset (Fig. 4.1c). It is generally accepted that substorms consist of both directly driven and loading-unloading processes. In this paper we consider nonlinear dynamical aspects of the global solar wind-magnetosphere interaction. We find that it is critically important to include magnetotail unloading in the model in order to replicate the main features of geomagnetic activity: With just the driven response in the model, we do not obtain realistic time behavior of the magnetosphere.

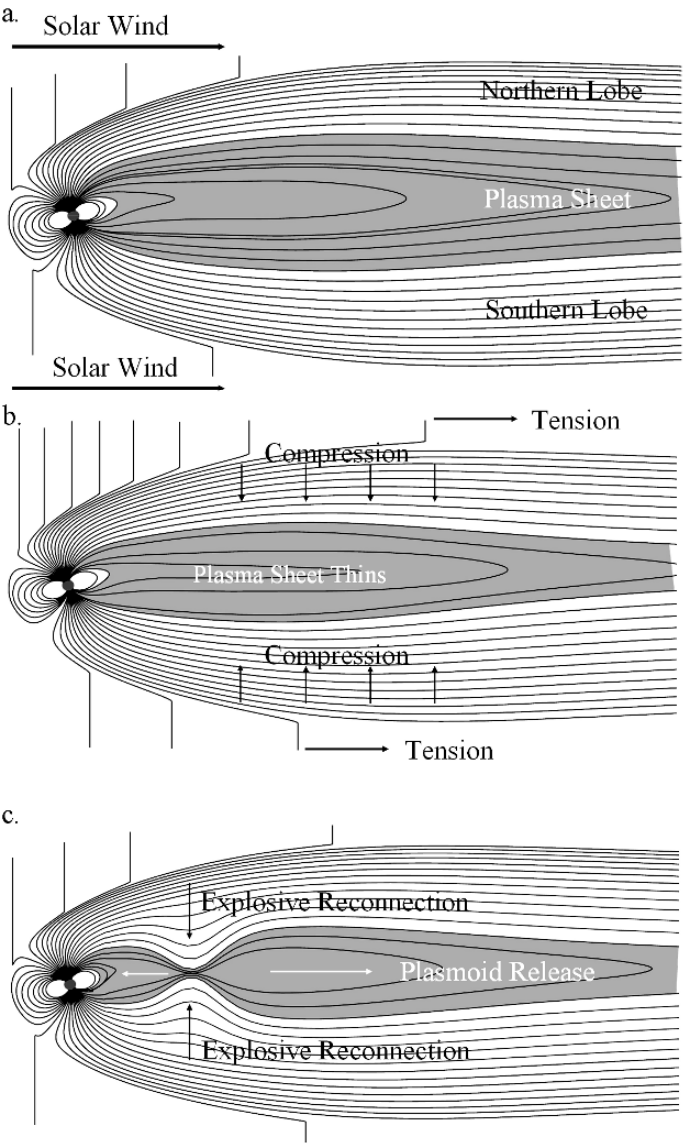


Fig. 4.1. Cross-sectional views of the Earth's magnetosphere showing the solar wind interaction and dynamical evolution. (a) The quiet time or "ground state" configuration showing weak solar wind energy coupling to the magnetosphere. (b) The growth phase of the magnetospheric substorm showing increased solar wind energy coupling (due to dayside magnetic reconnection) and a plasma sheet thinning on the nightside. (c) Explosive reconnection at a near-Earth magnetic neutral line in the plasma sheet and plasmoid formation at substorm onset.

The view of the solar wind-magnetosphere interaction we adapt here may essentially be termed the "standard" model (Baker et al. 1996). It is assumed that the energy coupling to the magnetosphere is effectively determined by the interplanetary dawn-dusk electric field, i.e., the product of the solar wind speed, V, and the southward component of the interplanetary magnetic field (IMF), B_s. A process of dayside merging at or near the subsolar magnetopause drives large-scale field-aligned currents over a large part of the polar regions and leads to transport of magnetic flux from the dayside to the magnetotail.

For weak dayside merging, the loading of flux into the magnetotail is quite modest and reconnection at the distant neutral line can balance the dayside merging rate. However, for strong B_s and/or high solar wind speed the dayside merging completely overwhelms the reconnection rate and return of flux from the distant X-line. This imbalance forces near-Earth neutral line development and substorm expansion phase onset (Fig. 4.1c). Reconnection in the near-tail rapidly returns magnetic flux to the dayside and rids the tail of large amounts of energy by release of the substorm plasmoid (Hones 1979; Baker et al. 1996). A major form of energy dissipation during substorms is the currents that flow from the magnetotail and close through the nightside ionosphere. These "unloading currents" produce high levels of Joule heating in the auroral zone near local midnight and can be monitored quite effectively by magnetic indices such as AE or AL.

2 Magnetospheric Substorm Dynamical Cycle

In an effort to study the relative roles of driven and unloading behavior during substorms, Bargatze et al. (1985) assembled an extensive solar wind and geomagnetic activity data set. They examined several dozen different intervals when a monitoring spacecraft was in the interplanetary medium during 1973-74. These intervals of complete, continuous solar wind coverage were typically several days long. The concurrent geomagnetic activity as measured by the AL geomagnetic index was then examined for these intervals. The 34 time periods selected for final inclusion in the data set were chosen so that geomagnetic activity for each of them went from quiet behavior ($AL \geq -50$ nT) through a more disturbed level and then back to quiet behavior. In effect, this complete quiet-to-quiet progression allowed the merging of the 34 separate intervals together as a single time series. The solar wind input parameter VB_s was then produced at 2.5-min resolution and was associated with the corresponding AL index (also 2.5-min resolution) for the same times.

In Fig. 4.2 we show the impulse response functions, or linear prediction filters, found by Bargatze et al. (1985) to describe the relationship between VB_s and AL. In this analysis, it was assumed that the solar wind input, I(t), is related to the magnetospheric output, O(t), by the convolution integral

$$O(t) = \int g(\tau)\, I\,(t-\tau)\, d\tau \qquad (1)$$

The linear prediction filter, g(t), gives the most general linear time-invariant relationship between I(t) and O(t). As noted above, Bargatze et al. used $I = VB_s$ and assumed

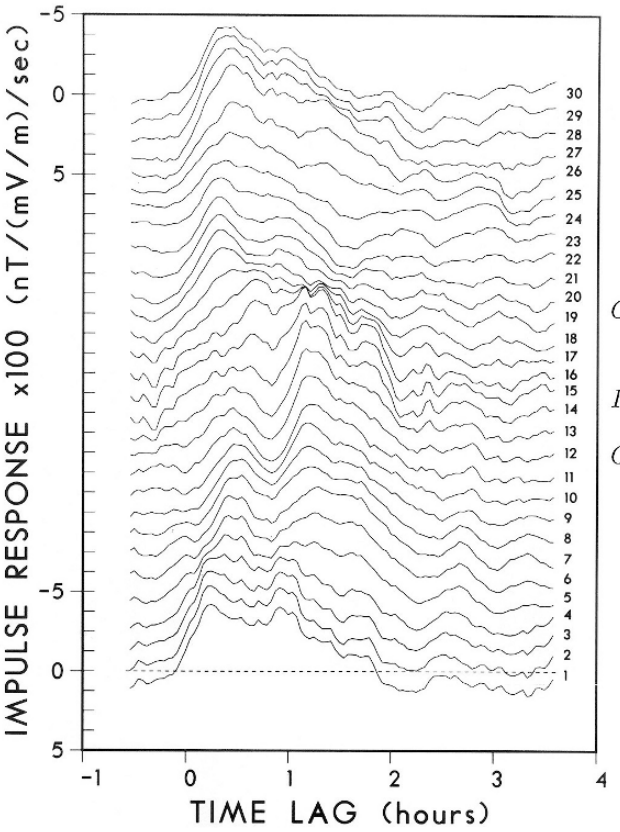


Fig. 4.2. A sequence of linear prediction filters (or impulse response curves) showing the time-lagged response between the interplanetary electric field (VB_s) and the AL geomagnetic activity index. The linear prediction filters are arrayed (bottom-to-top) from weakest average activity intervals to strongest activity. An evolution from double-peaked (bimodal) filters to single-peaked filters suggests a nonlinear dynamical behavior of magnetospheric activity (from Bargatze et al. 1985; reproduced by permission of American Geophysical Union).

O = AL. The curves in Fig. 4.2 (bottom to top) progress from weak through moderate to strong activity intervals.

It is seen in Fig. 4.2 that moderate activity filters have two distinct peaks as a function of time lag. The first, lower amplitude peak occurs at about 20-25 min lag while the second, stronger peak occurs at 60-70 min lag time. In contrast, the high-activity periods show one dominant peak at about 20-min lag with a broad shoulder extending to higher lag times. The interpretation of these results has been that the 20-min peak in the filters corresponds to directly-driven geomagnetic activity. It is due to dayside reconnection and global magnetospheric convection. It is manifested in all of the auroral electrojet indices (e.g., AL and AE), but is particularly evident in the eastward electrojet (as measured by the AU index). This 20-min timescale represents

the period it takes for the magnetosphere to readjust to new interplanetary conditions. Fundamentally this represents the large-scale current system associated with global transport of flux during the substorm growth phase. The 1-hour peak in Fig. 4.2 is taken as the time scale of tail unloading. It represents the delay time between an enhancement of VBs in the solar wind and the formation of the near-Earth neutral line. It corresponds, therefore, to the substorm plasmoid formation and to the tail current enhancement on the nightside of the Earth.

If the above interpretation of the linear filter elements in Fig. 4.2 is correct, then these results suggest that the magnetospheric response to solar wind input changes quite fundamentally as the system goes from weak to relatively strong driving conditions. For weak and moderate activity the magnetosphere exhibits both a driven and an unloading character whereas for strong activity the unloading response tends to be washed out. Baker et al. (1990) suggested that this change in the nature of the magnetospheric response represents an evolution toward essentially chaotic internal magnetospheric dynamics during strong solar wind driving conditions. This nonlinear evolution of magnetospheric dynamics was likened to the "dripping faucet" model that was first suggested by Hones (1979) in his plasmoid analogy shown here as Fig. 4.3.

3 Nonlinear Substorm Dynamical Behavior

Based upon the above interpretations and upon "phase-space reconstruction" methods, nonlinear dynamical models of the magnetosphere have been developed (Vassiliadis et al., 1995). These models attempt to replicate the internal magnetospheric dynamical evolution including loading of energy from the solar wind and dumping of energy from the system by plasmoid formation (e.g., Baker et al. 1990). The Faraday loop model of Klimas et al. (1992) in particular, provided a simple, but dynamically complete representation of the substorm cycle including energy storage in the magnetotail (growth phase), sudden unloading of tail energy (expansion phase), and a return toward the equilibrium or ground state (recovery phase).

Klimas et al. (1992) considered a closed loop in the magnetotail passing in the dawn-dusk direction through the cross-tail current sheet and then closing over the top of the tail lobe. They determined the total magnetic flux through this loop and calculated the changing flux through it according to Faraday's law: $1/c\ d\phi/dt = -\int d1 \bullet E$. With a current sheet width of d_{cs}, a cross-tail electric field (E_y) in the current sheet, and an externally imposed electric field (E) at the magnetopause, one has $1/c\ d\phi/dt = (E_o - E_y)\ d$. By estimating the lobe flux from the integral of $B \bullet dA$ and relating this to the cross-tail current density j_y, one can relate the time rate of change of flux to the area (A) of the lobe and the half-thickness (h) of the current sheet:

$$d/dt(hAj_y) = c^2/4\pi(E_o-E_y)d. \quad (2)$$

Klimas et al. then considered that the cross-tail current had three components, viz., a pressure balance term (j_o), a resistive term (j_R), and a polarization current (j_p).

Substituting plausible forms for each of the current components, the total current $j_y = j_o + j_R + j_p$ is expressed in terms of solar wind parameters, geometric quantities in the tail, and time variations of the cross-tail electric field. In dimensionless form, the nonlinear dynamical equation for the Faraday loop model is

$$d/d\tau[a(\alpha\beta + \nu E + dE/d\tau)] = (\varepsilon_o - E)\sqrt{a}. \quad (3)$$

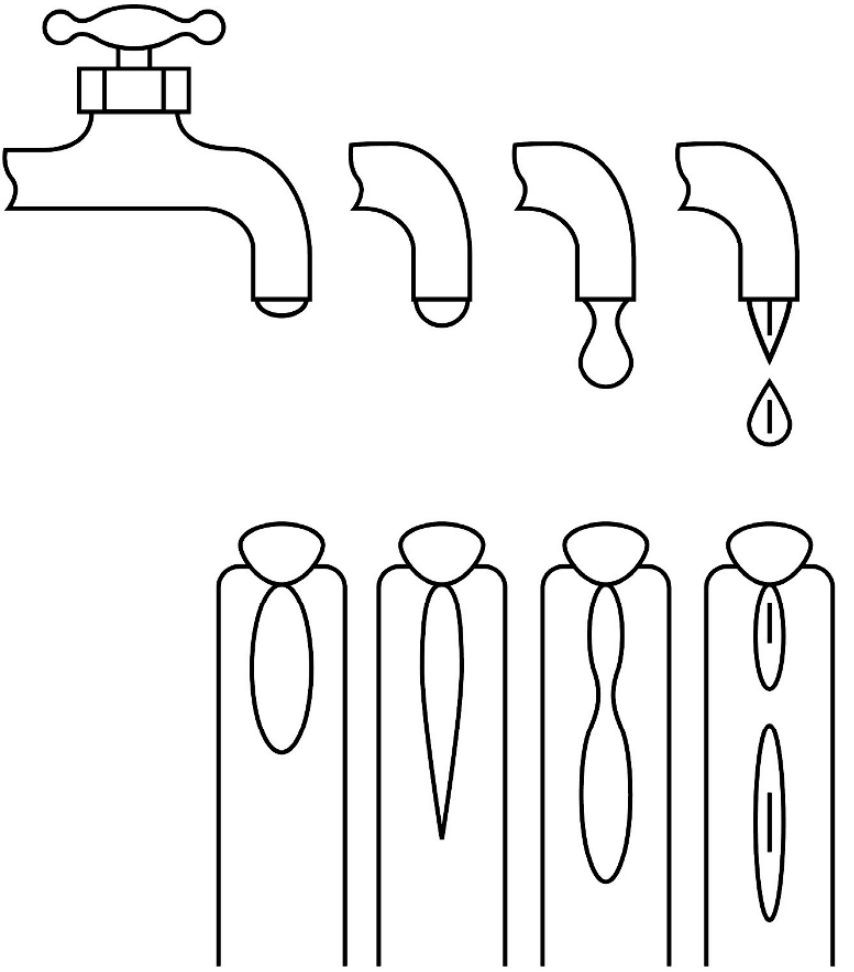


Fig. 4.3. An illustration of the analogy between a dripping faucet and the formation of a large-scale plasmoid in the Earth's magnetotail during a magnetospheric substorm (from Hones 1979). This concept has been extended to develop nonlinear analogue models of substorms (Baker et al. 1990; reproduced by permission of American Geophysical Union).

Here $\tau(=w\tau)$ is scaled according to the natural frequency of the system, a $(= (d/D)^2)$ is the dimensionless area of the tail lobe, and $\alpha(\propto \sqrt{D/h})$ is taken as a geometrical constant of the tail configuration. Note that D is the average diameter of the tail. Klimas et al. used the solar wind VxB electric field to scale E_o and E_y; thus, $\varepsilon_o \equiv E_y/E_{sw}$. Finally, the damping coefficient ν above is dependent on the plasma properties of the plasma sheet and the natural tail oscillation frequency, ω. The parameter β is the dimensionless tail lobe field strength necessary for pressure equilibrium. To close the equations, flux is loaded into the tail at a rate $\dot{\beta} = \dot{\beta}_L$ for loading conditions and $\dot{\beta} = \dot{\beta}_D$ for unloading. Thus $\dot{\beta}_D$ represents a much more rapid dissipation of the lobe flux which was added to the tail during the growth phase.

In prior work (Baker et al. 1993), the Faraday loop model was used and highly time-dependent solar wind driving of the model was considered. The observed VB_s from the Bargatze et al. (1985) data set was used as a driver for the Faraday loop model. The result of such work showed that the measured values of AL compare well with the values of AL computed from the Faraday loop model using VB_s as the driver. Without unloading in the model, one can fit reasonably well the low frequency response of the magnetosphere to the solar wind, but one cannot very well fit the sharp substorm onsets or the large amplitude changes in AL: These discrete features are the essence of substorms. Thus the analogue model with realistic solar wind driving shows results very similar to prior studies of the real magnetosphere, namely, that with both driven and unloading aspects included, the analogue model provides a reasonably good replication of the observed AL time series.

4 Multi-scale Aspects: Self-Organization

Baker et al. (1999) considered the nature of the very onset of magnetospheric substorms. As shown in Fig. 4.4, it was recognized that the magnetotail has essentially two dominant states: (a) a highly stressed configuration; and (b) a relaxed configuration. The substorm onset can be viewed as a catastrophic transition from the stressed state to the relaxed state. Much of the continuing debate in magnetospheric substorm research today revolves around the nature and location of the instabilities in the magnetosphere that lead to the catastrophic transitions depicted in Fig. 4.4 (see also, Lewis 1991). As noted by Baker et al. (1999), there may be several plasma physical processes that lead to relaxation during substorms and different instabilities may dominate during different events.

Many researchers have tried to develop simple (or at least simplified) models of the complex, nonlinear interactions that characterize the real terrestrial magnetosphere. Many authors (e.g., Chapman et al.,1998; Chang,1999) have utilized rather basic models to try to understand the relationship between small-scale phenomena in the Earth's plasma sheet and ionosphere and the global-scale reconfiguration of substorms. Chapman et al. (1998), in particular, used "sand pile" models that are

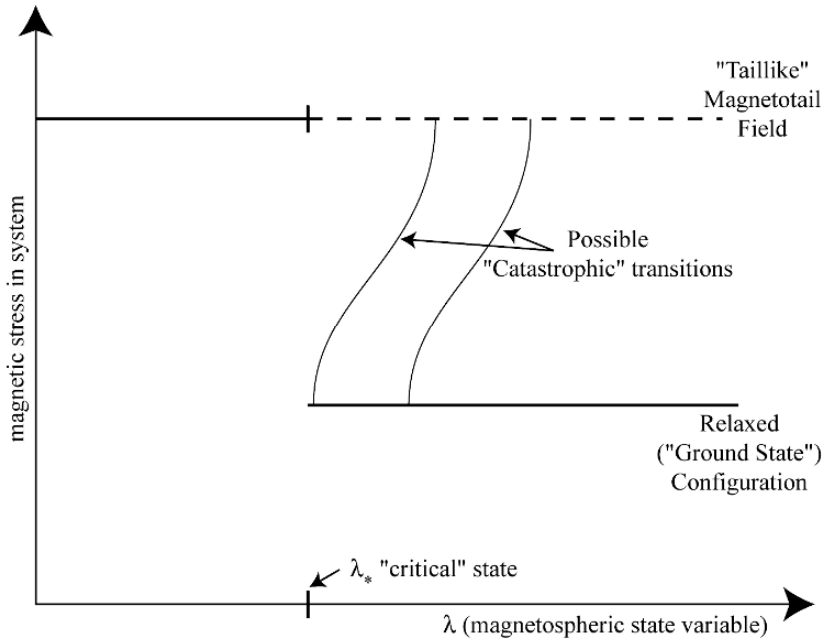


Fig. 4.4. A diagram illustrating the abrupt ("catastrophic") transition of the magnetosphere from a highly stressed, or taillike, magnetic configuration to a much more relaxed configuration. Several different physical pathways for such relaxation may be possible once the magnetosphere exceeds a critical value of a magnetospheric state variable (λ_*) (adapted from Baker et al. 1999).

prescribed cellular automata models to describe slow build up and avalanching releases as an analogue of the magnetotail and substorms.

From observed dynamics and reflections on the complexity of the magnetospheric systems, we can conclude the following:

- The solar wind-magnetosphere-ionosphere system clearly is characterized by nonlinear dynamics
- Methods borrowed from other branches of physics, chemistry, and engineering can offer useful – if imperfect – analogies
- Ultimately, space plasma physics must depart from idealized local "stability" analyses to consider true global system responses

These ideas present strong motivations to go away from usual plasma physics techniques to consider completely different approaches. It may very well be that the magnetotail has a distributed set of interacting control agents, some of which are local plasma conditions and some of which are remote in the ionosphere or solar wind. It may, therefore, appear that instabilities are spontaneous or even random (e.g. Lewis 1991). We must use new analysis tools to address such issues. Some of the most innovative results in data analysis pertaining to magnetospheric dynamics have come about recently in the analysis of auroral imaging data. Several papers have attempted to address the questions of coupling and multi-scale phenomena through the analysis of spatial and temporal scales contained in global auroral images (Lui 2002; Uritsky et al. 2002). Here we describe some of these results and their interpretation.

Figure 4.5 shows an example of ultraviolet images (top panel) from the Ultraviolet Imager (UVI) experiment on the POLAR spacecraft. The lower panel shows that the auroral brightening in the upper panel – and its temporal evolution – is well associated with a strong tailward plasma flow event measured by the GEOTAIL spacecraft in the Earth's nightside plasma sheet. Magnetic fluctuations on many temporal (and spatial) scales are seen in the magnetic field data from GEOTAIL (lower panel) for this event.

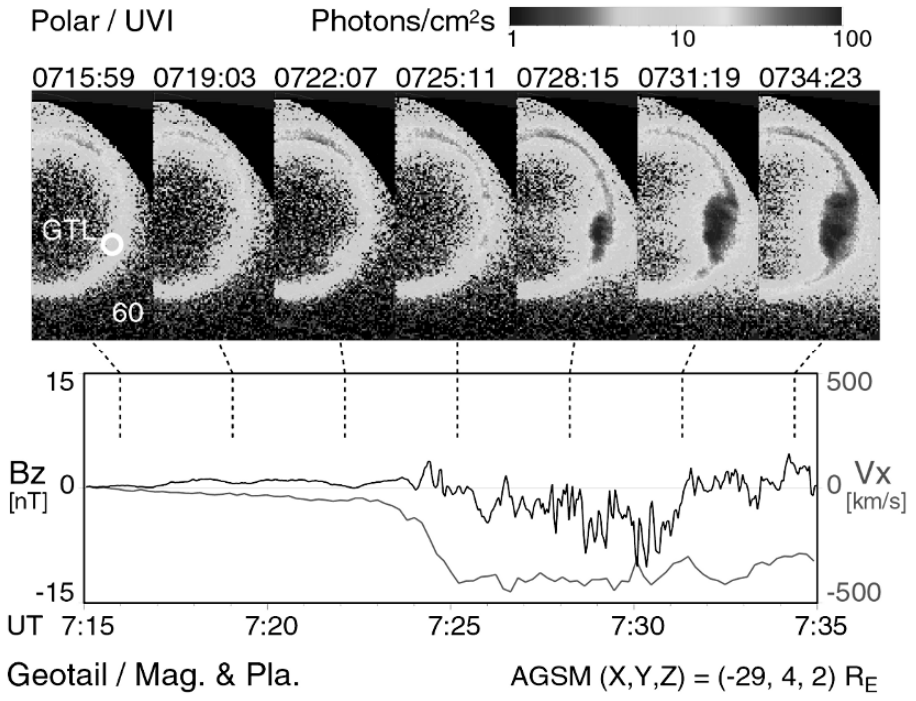


Fig. 4.5. POLAR Ultraviolet Imager (UVI) images of the northern auroral oval showing a strong substorm onset event at ~0725 UT (upper panel). This onset is well-associated with strong tailward plasma flow and magnetic field signatures (lower panel) seen concurrently by the GEOTAIL spacecraft on the nightside of the Earth (Ieda et al., 2001). (A color version of this figure appears between pages 196 and 197).

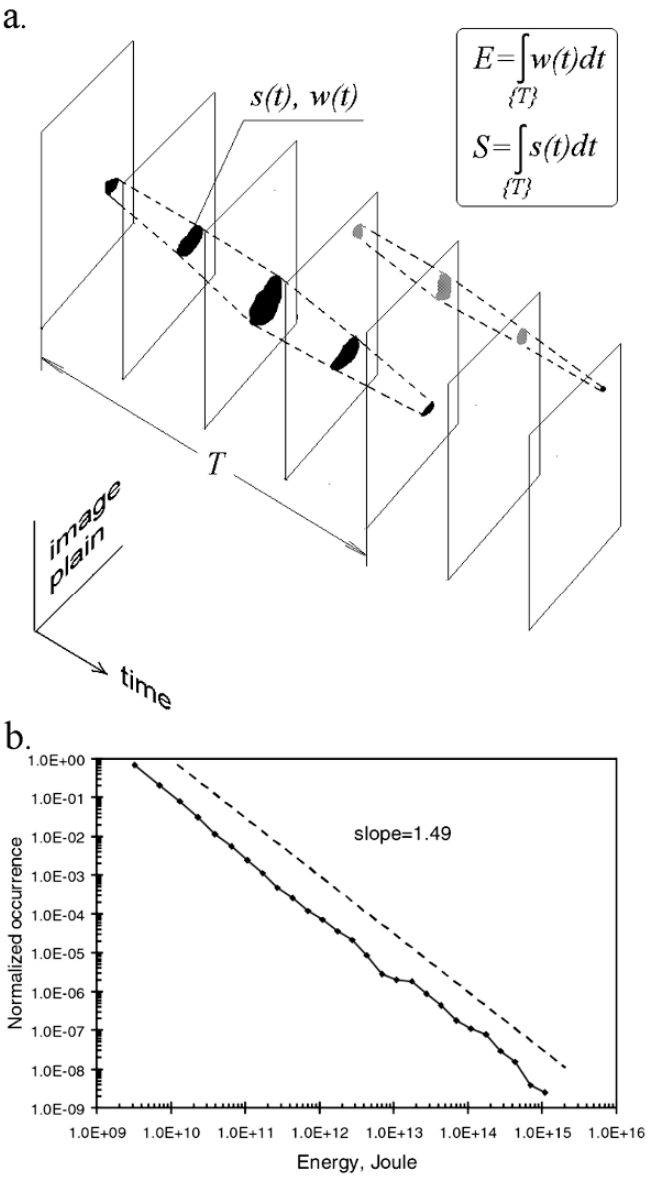


Fig. 4.6. (a) A schematic illustration of the method used to identify and track the spatial and temporal auroral features in the Ultraviolet Imager (UVI) data set from the NASA/POLAR spacecraft. (b) Normalized occurrence frequency distributions versus integrated energy content determined for POLAR/UVI images during 1997 and 1998 (from Uritsky et al. 2002; reproduced by permission of American Geophysical Union). A clear power law (scale-free) distribution is seen.

As shown in Fig. 4.6a, Uritsky et al. (2002) used a very innovative method of analyzing data from the POLAR spacecraft (of the sort shown in Fig. 4.5). Unlike earlier work, Uritsky et al. followed a given auroral feature from image to image over the entire time of its existence in the UVI sequences. Thus, they were able to integrate a given auroral feature both as to its spatial extent and also as to its temporal duration. From the standpoint of total energy dissipation in a given auroral feature, such tracking of events across the entire time of their persistence is critical.

Figure 4.6b shows an example of the results of the Uritsky et al. (2002) spatiotemporal statistical analysis. The authors studied ionospheric emissions seen in the UVI experiment for four months in 1997 and 1998. There were some 12,300 individual emission events tracked in space and time. Figure 4.6b shows the occurrence frequency distribution in terms of the estimated total energy content of each identified emission event. A nearly perfect power law distribution (in this case going as $E^{-1.49}$) was found. Uritsky et al. noted that five or more orders of magnitude in scale size are described by these power law fits. These results clearly suggest a nearly "scale free" behavior of auroral enhancements and related magnetotail (plasma sheet) dissipation events.

As argued by Uritsky et al. (2002) and Klimas et al. (2005), their statistical results strongly imply that the observed behavior is consistent with that of an avalanching system near a (stationary) critical state. Hence, these results support the notion that the magnetotail exhibits self-organized critical dynamics (see also, Chang 1999). Since substorm luminosity events are fully within the POLAR/UVI data set of Uritsky et al. (2002), these results suggest that there is a remarkable occurrence of cross-scale coupling effects from the smallest events in the plasma sheet (and ionosphere) to the largest spatial and temporal scales of global substorms.

5 Future Directions: Modeling and Missions

The kinds of power law distributions found by Uritsky et al. (2002) shown here in Fig. 4.6 are very suggestive of the "scale-free" distributions that are expected to result from MHD turbulence (Kolmogorov 1941). Such models describe an energy cascade process from large to small scales. Kolmogorov turbulence is generally expected to follow a wave number spectrum going as $k^{-5/3}$. Relatively recently other authors (e.g., Goldreich and Sridhar, 1995) have investigated MHD turbulence in terms of nonlinear interactions among shear Alfven waves. It is interesting that the energy scale distribution found for auroral dissipation events (Fig. 4.6) follows a spectral distribution that is close ($\sim E^{-1.5}$) to that which might be expected to result from turbulent eddies that exist in the magnetically-conjugage plasma sheet.

In their brief review, Sharma et al. (2005) considered the possible relationships between phase transitions, self-organization, and plasma turbulence as these might bear on magnetospheric dynamics and substorms. Sharma et al. noted that each of these (seemingly) disparate constructs had some attractive aspects in explaining magnetospheric temporal and spatial behavior. However, no one approach seems fully to explain the complex evolution of magnetospheric properties during strong dynamical events.

In recent modeling work, Klimas et al. (2000, 2005) have proposed that localized magnetic reconnection in the turbulent plasma sheet in Earth's magnetotail could evolve in a collective way to form an energy avalanche. This could then lead to the kind of scale-free behavior that has been interpreted as self-organized criticality in auroral features, etc. Klimas et al. (2005) have developed a set of numerical simulations of magnetic field annihilation (using resistive MHD codes) to study plasma sheet dissipation. Present work using 2-D models (but not yet 3-D codes) suggests that localized magnetic reconnection starting on very small spatial scales can rapidly evolve into self-organized criticality on more global scales. Further self-consistent modeling of this sort is crucial for tying together various views of magnetospheric dynamics.

Ultimately, it will be necessary to have better multi-scale observations of plasma sheet properties in order to confirm (or refute) numerical simulation results. As noted by Sharma and Curtis (2005), new spacecraft missions are now on the horizon that

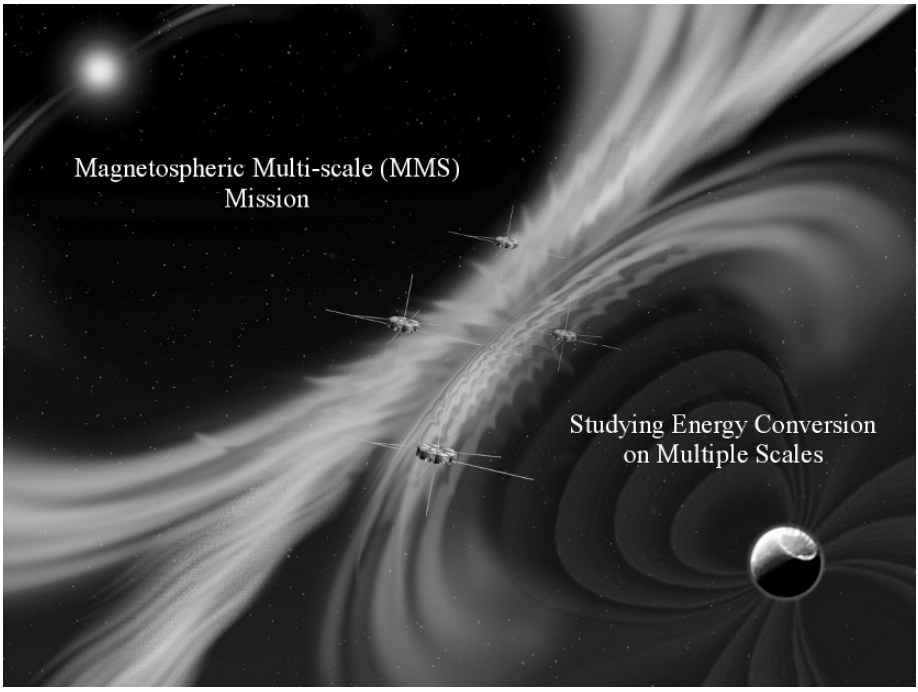


Fig. 4.7. An illustration of the NASA Magnetospheric Multi-scale (MMS) four-satellite constellation of observatories now approved for launch in 2013. The four MMS spacecraft will be able to be configured in a tetrahedral flight pattern with satellite separations ranging from a few hundred meters to thousands of kilometers (courtesy of J. Burch). MMS will allow examination of energy dissipation and magnetic reconnection in Earth's magnetosphere on a vast range of spatial scales. (A color version of this figure appears between pages 196 and 197).

will have great potential to explore magnetic field, plasma, and energetic particle phenomena in the Earth's magnetosphere from the microscale to the macroscale. As shown here in Fig. 4.7, the Magnetospheric MultiScale (MMS) mission of NASA will consist of four identical spacecraft flying in a tight tetrahedral configuration. MMS will focus on the boundary layers between various magnetospheric regions, but it will also provide detailed spatio-temporal information of reconnection and energy dissipation events throughout the Earth's plasma sheet. Using the comprehensive measurement payload onboard MMS (J. Burch, P.I.), MMS will be able to study thin current sheets, plasma turbulence, and particle acceleration on scales ranging from the electron gyro-radius up to scales of substorm plasmoids themselves.

6 Discussion

Scientific research often proceeds by separating a system into simple component parts, and explaining each part individually. Such an approach leads to descriptions of complicated systems in terms of a large number of variables, each looked at separately. Each part has the convenient property of linearity, meaning that output is proportional to input. This strategy, however, does not always work well. Scientists and engineers have known for many years that systems can be intrinsically nonlinear, having output that is not proportional to input. Some nonlinear systems exhibit feedback, in which output influences input. Feedback can be stable, as with the thermostat system that helps control room temperature. It can also be unstable, as in lightning, explosions, or wave phenomena, in which the feedback processes magnify conditions giving rise to the effect. Nonlinear systems can exhibit seemingly random behavior, as in the turbulent swirling of a stirred fluid, or in weather patterns. Such "deterministic chaos" shows simple nonlinear dynamical systems that exhibit randomness. It is even possible to demonstrate chaos in systems with only three variables, if the parameters and couplings are right (e.g., Klimas et al. 1996).

Nonlinearity can also give rise to the phenomenon of self-organization, which can be thought of as order emerging from chaos. Nonlinearity can cause randomness and order in a single system with small or large numbers of variables, and this is one form of complexity. Another situation that promotes complexity is the coupling of several systems, even ones that are described by entirely different variables. Such system level complexity occurs in economics, in which many different types of systems (manufacturing, finance, wages, consumer spending, etc), interact strongly with one another, or in biological organisms in which many systems interact to maintain homeostasis. There are examples resulting from nonlinear couplings in a single system, and there are examples resulting from coupling to two or more different systems. Generally what one means by a system is an entity described by a particular set of conditions, a particular set of boundaries and a particular set of mathematical equations. In this sense, the wind, the sand and the waves of the ocean are each a nonlinear system, and, in addition, they are strongly coupled to one another as well. So too, are the solar wind, the magnetosphere, the ionosphere, and the neutral atmosphere. Each is complex and

nonlinear in its own right, but they are also strongly influencing one another. Thus, many types of complexity commonly occur in solar and space physics.

The idea of self-organization has played a very strong role in many branches of modern science. For example, in a recent book review, it was stated that:

"The concept of self-organization originated in the study of nonlinear physical and chemical systems, such as convection flows and chemical reactions that form waves. In these systems, global patterns emerge from local interactions among many subunits. The interactions are typically shaped by multiple feedback loops, including positive ones that amplify emergent dynamics and negative ones that modulate and constrain them." (Pepper and Hoelzer 2001).

This statement could very easily be made about, and be applied to, the magnetosphere-ionosphere-atmosphere system. As shown in this brief review paper, several aspects of recent substorm research point to global patterns in substorm dynamics that must be intimately related to smaller scale features in the magnetosphere. Continuing modeling and new spacecraft missions hold real potential for resolving the outstanding issues of nonlinearity and complexity in magnetospheric dynamics.

References

Baker, D.N., Klimas, A.J., McPherron, R.L., and Buchner, J. (1990) The evolution from weak to strong geomagnetic activity: An interpretation in terms of deterministic chaos, Geophys. Res. Letters, 17, 41.

Baker, D.N., Klimas, A.J., Pulkkinen, T.I., and McPherron, R.L. (1993) Re-examination of driven and unloading aspects of magnetospheric substorms, Adv. Space Res., 13, #4, 475-483.

Baker, D.N., Pulkkinen, T.I., Angelopoulos, V. , Baumjohann, W., and McPherron, R.L. (1996) The neutral line model of substorms: Past results and present view, J. Geophys. Res., 101, 12,995-13,010.

Baker, D.N., Pulkkinen, T.I., Büchner, J., and Klimas, A.J. (1999) Substorms: A global instability of the magnetosphere-ionosphere system, J. Geophys. Res., 104, A7, 14,601-14,611.

Bargatze, L.F., D.N. Baker, R.L. McPherron, and E.W. Hones, Jr. (1985) Magnetospheric impulse response for many levels of geomagnetic activity, J. Geophys. Res., 90, 6387.

Chang, T. (1999) Self-organized criticality, multi-fractal spectra, sporadic localized reconnections and intermittent turbulence in the magnetotail, Phys. Plasma, 6(11), 4137-4145,

Chapman, S.C., Watkins, N.W., Dendy, R.O., Helander, P., and Rowlands, G. (1998) A simple avalanche model as an analogue for magnetospheric activity, Geophys. Res. Lett., 25, 2397.

Goldreich, P., and Sridhar, S. (1995) Toward a theory of interstellar turbulence. 2. Strong Alfvenic turbulence, Astrophys. J., 438, p. 763.

Hones, E.W., Jr. (1979) Transient phenomena in the magnetotail and their relation to substorms, Space Sci. Rev., 23, #3, 393-410, doi: 10.1007/BF00172247.

Klimas, A.J., Baker, D.N., Roberts, D.A., Fairfield, D.H., and Buchner, J. (1992) A nonlinear dynamic analogue model of geomagnetic activity, J. Geophys. Res., 97, 12253.

Klimas, A.J., Uritsky, V.M., Vassiliadis, D., and Baker, D.N. (2005) Simulation study of SOC dynamics in driven current-sheet models. In: A.S. Sharma and P.K. Kaw (eds*.) Nonequilbrium Phenomena in Plasmas*, pp. 71-89.

Klimas, A.J., Valdivia, J.A., Vassiliadis, D., Baker, D.N., Hesse, M., and Takalo, J. (2000) Self-organized criticality in the substorm phenomenon and its relation to localized reconnection in the magnetospheric plasma sheet, J. Geophys. Res., 105, 18,765.

Klimas, A.J., Vassiliadis, D., Baker, D.N., and Roberts, D.A. (1996) The organized nonlinear dynamics of the magnetosphere, J. Geophys. Res., 101, A6, 13,089-13,113.

Kolmogorov, A.N. (1941) The local structure of turbulence in incompressible viscous fluids at very large Reynolds numbers, Dokl. Akad. Nauk. SSSR, 30, 301-305. Reprinted in 1991: Proc. R. Soc. Lond. A, 434, 9-13.

Lewis, Z.V. (1991) On the apparent randomness of substorm onset, Geophys. Res. Lett., 18, 1849.

Lui, A.T.Y. (2002) Multiscale phenomena in the near-Earth magnetosphere, J. Atmos. and Solar-Terr. Phys., 64, 125-143.

Pepper, J. W., and Hoelzer, G. (2001) Unveiling Mechanisms of Collective Behavior, Science, 294.

Sharma, A.S. and Curtis, S.A. (2005) Magnetospheric Multiscale Mission, In: A.S. Sharma and P.K. Kaw (eds*.) Nonequilbrium Phenomena in Plasmas*, pp. 179-196.

Sharma, A.L., Baker, D.N., and Borovsky, J.E. (2005) Nonequilbrium phenomena in the magnetosphere: Phase Transition, self-organized criticality, and turbulence, In: A.S. Sharma and P.K. Kaw (eds*.) Nonequilbrium Phenomena in Plasmas*, pp. 3-22.

Uritsky, V.M., Klimas, A.J., Vassiliadis, D., Chua, D., and Parks, G. (2002) Scale-free statistics of spatiotemporal auroral emissions as depicted by POLAR UVI images: The dynamic magnetosphere is an avalanching system, J. Geophys. Res., 107, 1426.

Vassiliadis, D., A.J. Klimas, D.N. Baker, and D.A. Roberts, A description of the solar wind-magnetosphere coupling based on nonlinear prediction filters, J. Geophys. Res., 100, 3495-3512, 1995.

5 Microseism Activity and Equilibrium Fluctuations

Antoni M. Correig[1,2], Mercè Urquizu[2], Josep Vila[2,1] and Ramon Macià[3,2]

[1] Universitat de Barcelona, Departament d'Astronomia i Meteorologia, Ton.Correig@am.ub.es
[2] Institut d'Estudis Catalans, Laboratori d'Estudis Geofísics "Eduard Fontserè", merce_urquizu@yahoo.es, jvila@am.ub.es
[3] Universitat Politècnica de Catalunya, Departament de Matemàtica aplicada II, ramon.macia@upc.edu

Abstract. A study is carried out on the seismic wave field recorded in the absence of earthquakes, the unrest of the solid earth, in the frequency band of microseism activity. Two phenomenological models are presented to explain the main characteristics of the observed time series from the point of view of both the power spectrum and the phase space representation, corresponding to the cases of strong external forces (well developed microseism activity) and very low ones (minimum energy spectra, termed equilibrium fluctuations). Observations and model predictions of equilibrium fluctuations strongly suggest that the information contained in the power spectrum corresponds to medium resonances, that may strongly fluctuate, through a competition process due to the action of external forces.

1 Introduction

It is well known for more than a century that the Earth, in the absence of earthquakes, undergoes continuous oscillations (Wiechert 1904; Rhie and Romanovicz 2004), in all frequency range, from 0.002 mHz to 20 Hz, with different amplitudes. In the high frequency interval (1 Hz – 20 Hz) these oscillations, known as local and cultural noise (due to local meteorological conditions and industrial activities) are considered to be random (Lachet and Bard 1994), whereas the low frequency range (0.002 Hz – 0.01 Hz) respond to a linear deterministic dynamics (Rhie and Romanovicz 2004), because the observed eigenfrequencies are well predicted by the linear theory of free oscillations. The aim of this study is to explore the interval range 0.04 Hz – 0.3 Hz, the domain of microseism activity (Webb 1998); at lower frequencies, between 0.01 Hz – 0.04 Hz there is the influence of infragravity oceanic waves, and at higher frequencies, between 0.3 Hz – 1 Hz, there is the influence (resonances) of regional and local structures (Rial 1989).

Microseism activity is very often considered as *noise* because it masks the arrival of coherent seismic phases. From the early days it was recognized that the source of well developed microseism activity is atmospheric storms at sea that generate quasi-stationary perturbations propagating as normal modes, mainly as Rayleigh waves

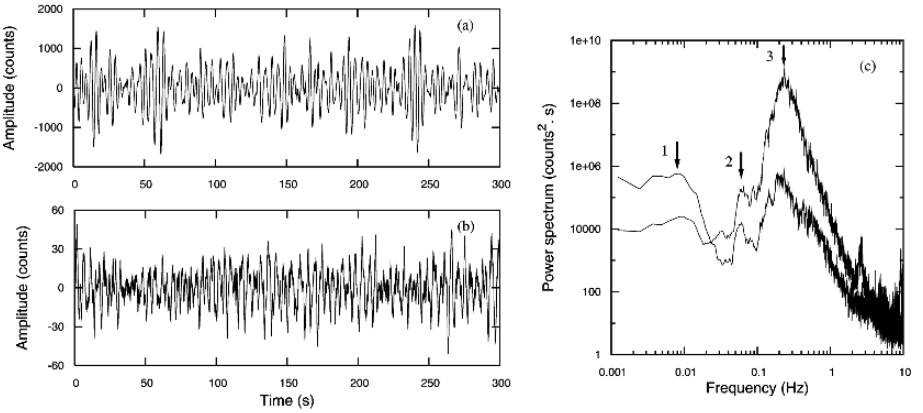


Fig. 5.1. Examples of two microseism time series recorded at CAD station corresponding to a stormy day, (a), with large amplitude, and to a calm day (b) of low amplitude; note the difference in amplitude scales. (c) displays the power spectrum of the time series (a) and (b); apart of the difference in amplitudes, both spectra share the same characteristics: a peak located at ~0.016 Hz corresponding to infragravity waves, the primary peak located ~0.07 Hz and the secondary peak located at ~0.2 Hz.

(Press and Ewing 1948). Two main mechanisms of microseismic generation have been proposed: the surf mechanism (Wiechert 1904) generating primary microseisms with a period equal to the storm wave period (a frequency of about 0.07 Hz), and the interference mechanism (Longet-Higgins 1950), generating secondary microseisms as the result of fluctuations of pressure caused by standing waves along the sea bed, with a period equal to half of that of the storm waves (a frequency of about 0.14 Hz but with a large scattering to higher frequencies). Both mechanisms, complementary, predict an approximate frequency relation of 2:1, although with a large observational scattering. As a general rule, spectral characteristics of microseismic activity are interpreted in terms of source mechanism (Longet-Higgins 1950), irrespective of medium response (Hasselman 1963), implicitly assumed to be "transparent" or able to be accounted for.

Figure 5.1 displays two examples of microseism time series recorded by the broadband seismic station CAD located at the eastern Pyrenees (Vila 1998), corresponding to a situation of high storm activity, top, and to a situation of a very low activity, bottom (see the amplitude scale). In the frequency domain the main features of microseism time series correspond, see fig. 5.1 (c), to a broad spectral peak close to 0.016 Hz, the *infragravity peak*, another centered at 0.07 Hz, the *primary peak*, and another centered at 0.2 Hz, the *secondary peak*. Apart of the amplitude, the records in both time and frequency domains are quite similar. It is interesting to note that two different wave-packets can be clearly identified in time domain, see Fig. 5.2: one of length 15 – 17 s (related to the primary peak), and another of length several tens of seconds (related to the infragravity peak), both composed of an oscillation of a period of ~5 s (related to the secondary peak). These wave packets, often masked by transients, deserve further study.

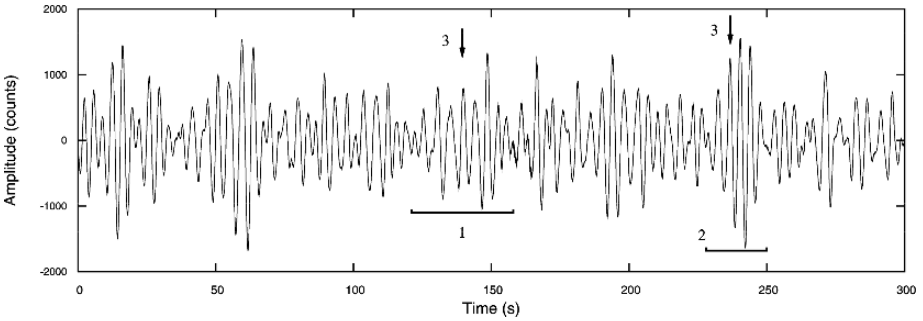


Fig. 5.2. Interpretation of the spectral peaks of Fig. 1 (c) in terms of wave packets for the case (for clarity) of the large amplitude time series. Two wave packets can be clearly distinguished of length several tens of seconds and 15– 20 seconds, both composed of quite stable oscillations of ~5 s (peak of 0.2 Hz). These wave packets suggest a possible interpretation in terms of solitons.

A lot of theoretical studies, using linear theory, have been carried out to explain these spectral characteristics, especially for oceanic records in which the action of the wind and waves acting as external forces is important, see (Webb 1998) for a review.

(Vila 1998) and (Vila and Macià 2002), in a systematic study of microseism activity recorded by a broadband seismic station, found a very well defined periodic variations of the amplitude of the spectral peaks well correlated to climatic seasonal variations, as well as strong fluctuations in the location and amplitude of the spectral peaks apparently well correlated to local and regional meteorological conditions. The (*FDSN Station book* 1994) present a world wide average (annual) values of broadband microseism spectra, all displaying similar characteristics of those of Fig. 5.1 (c), but with a frequency relation of primary to secondary peaks closer to 3:1 than 2:1, the same as in Fig. 5.1 (c).

(Correig and Urquizú 2002) and (Ryabov, Correig, Urquizú, and Zaikin 2003) performed an exhaustive non linear analysis of microseism time series and found that they are non-stationary, non-linear and stochastic. The *non-stationarity* has been accounted for in two different ways: *i)* because of the no conservation of the two first moments, that is, the mean and the variance for a minimum interval of the time series of 300 s (the microseism spectral peak is located at ~0.2 Hz, corresponding to a period of ~5 s) and *ii)*, by applying the stationary test provided by Bendat and Piersol 1986, p. 342). A possible explanation for the non-stationarity of the recorded time series may relay in that the quasi harmonic external forces may be close to the resonant response of the medium; numerical simulations with our model support this hypothesis.

The *stochasticity* has been inferred because of the non-saturation of the correlation dimension for an embedding dimension of phase space up to 15; it could be due to a noise source component. The *non-linearity* is inferred because otherwise it is not possible to simulate the evolution of the system in phase space, and can be due to a nonlinear source process, non-linear medium or both.

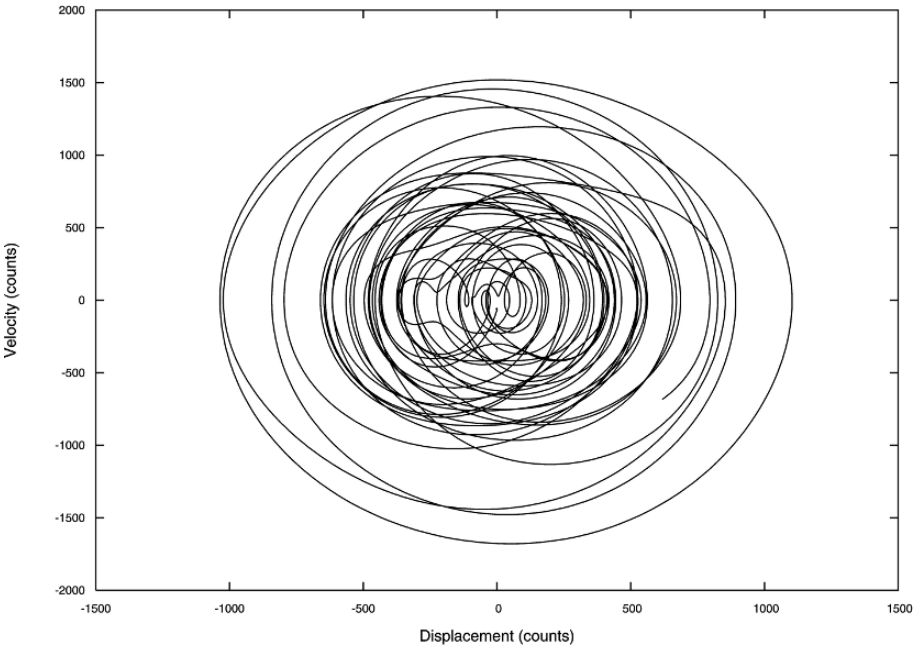


Fig. 5.3. Motion in phase space of the time series of large amplitude displayed in Fig. 5.1 (top) along with a zoomed view (bottom). The motion is clearly chaotic, with never repeating paths.

Figure 5.3 displays the evolution of microseism time series in phase space displacement *vs*. velocity (as broadband stations is a velocity record, the displacement have been computed by numerical integration). As it can be seen, the motion follows well-defined trajectories, similar to those of a particle bouncing irregularly in a potential well, consisting of a superposition of loops of different mean radius (*i.e.* motion with different frequencies) with the center of the loops displaying separate irregular oscillations, over a well-defined path. The corresponding motion is random in the sense that it is not possible to predict either the time evolution of the center of the loops or its mean radius. In other words, we are dealing with a chaotic system in the presence of noise, well differentiated from a random system (found in the high frequency range) and from a deterministic linear system (found in the low frequency range).

The aim of this study is to get some insight, through phenomenological models, on the underlying mechanism of unrest at this intermediate frequency range.

This study is an integrated version of two previous papers, (Correig and Urquizú 2002) and (Correig, Urquizú, Macià, and Vila 2006), on which it relies, and it attempts to present a clear discrimination between *i)* the well developed microseism activity that propagates mainly as coherent Rayleigh waves with a well defined source location, and *ii)* the underlying equilibrium fluctuations, a manifestation of the medium resonance of the Earth for this period range. As a main result, we have found that the spectral characteristics of microseism activity consist on fluctuations

(shifts) of the main characteristics of the fluctuation equilibrium due to a competition between the external atmospheric/oceanic forces and the resonant medium response, contrary to the commonly accepted hypothesis that microseism spectral peaks are a direct manifestation of the source activity

Quoting (Bath 1973) we can say that "the studies of microseisms, the steady unrest of the ground, are a border-line field between meteorology, oceanography and seismology. Microseisms are no doubt a greater concern to seismologists, but when their generation is to be explained, recourse must be taken to meteorological and oceanographic conditions. As a consequence, microseism constitutes a random process, like atmospheric turbulence and ocean surface waves". To accomplish this goal we will make use of two working hypothesis: *i)* the model has to be constructed taking into account the information provided by both the spectral analysis and the motion in phase space (as the time series is not stationary, it has no meaning to work in time domain) and ii), the main (global) characteristics of the system will be uncovered from observations at the lowest energy contents, to get rid of (local) spurious transients such as the occurrence of earthquakes or atmospheric storms.

2 Phenomenological Modeling of Microseism Time Series

Our aim is to develop phenomenological models, able to explain the observations, taking into account the two working hypothesis advanced in the Introduction. As a lot of studies have been carried out taking into account only the information provided by the power spectrum, we will emphasize the information provided by the evolution of the system in phase space, clearly displaying a chaotic behavior, and hence nonlinearity. Moreover, we will be concerned wit the signals recorded at a fixed point, that is, the oscillatory motion at the recording site, not with wave propagation, although both points of view are related. Effectively, as shown by (Xie and Gao 2004), traveling-wave solutions of nonlinear partial differential equations u(x,t) (as we will hypothesize later on) of the kind

$$u_{tt} + au_{xx} + bu + cu^3 = 0 \tag{1}$$

can be transformed into nonlinear ordinary differential equation for $u(\xi)$ under the change of variable $\xi = k(x - \lambda t)$, where k and λ are constants. Hence, we can simplify our original problem dealing with traveling waves to the analysis and simulation of oscillatory motions. As a working hypothesis, we postulate that our observations can be simulated by a nonlinear damped forced oscillator.

As commented at the introduction, two main mechanisms have been proposed for the origin of microseism activity, surf (Wiechert 1904) and interference (Longet-Higgins 1950) that in a first approximation can be considered as periodic or quasiperiodic. The same can be said for the infragravity waves. We thus approximate these contributions as external periodic forces. As the high frequency contents of the spectrum appear to be random, we also add white noise as a source term. To account for the nonlinearity, a monopotential Duffing oscillator has been accepted, as suggested by Fig. 5.3, and to account for all characteristics of the observed variability of the spectra a parametric resonance has been added. Note that the model has been derived with the aim to account for observations, no from first principles. Hence, we are dealing with a phenomenological or *toy* model. The model, a nonlinear damped forced oscillator reads

$$\begin{aligned} &\frac{d^2q}{dt^2} + \frac{dV}{dq} + \delta\frac{dq}{dt} = \sum_{i=1}^{2}\gamma_i \cos(\omega_i t) + \varepsilon F(t) \\ &V(q) = -\alpha\frac{q^2}{2} + \beta\frac{q^4}{4} \\ &\alpha = \alpha_0 + \eta\, f(t) \end{aligned} \tag{2}$$

where q stands for displacement, V stands for the nonlinear potential (monostable for $\alpha < 0$ and bistable for $\alpha > 0$), δ is the damping coefficient, γ_i are the amplitudes of the harmonic external forces, ε is the amplitude of the random fluctuations $F(t)$, α is the term of parametric resonance and η the amplitude of the time function $f(t)$ (sinusoidal or random). Fig. 5.4 displays an example of the evolution in phase space of a time series generated by eq. (2) with only one potential well, with the following values of the parameters: $\delta = 0.01$, $\beta = 0.05$, the frequencies of the harmonic external forces $f_1 = 0.05$, $\gamma_1 = 7.5$, $f_2 = 0.2$, $\gamma_2 = 1.0$, $\varepsilon = 10.0$, $\alpha_0 = -4.0$ and $\eta = 0.03$ with $F(t)$ white noise. We can see that qualitatively, the agreement with Figs. 5.3 and 5.4 is quite good. Let us emphasize that whereas in a linear system the invariants are the resonances, and in a chaotic system the invariants are the dimension of the attractor and the Lyapunov exponents, for the case of a stochastic, nonlinear and non stationary dynamical system we are not aware (up to the present) of the existence of invariants to quantify a state, so the need of qualitative comparisons. In our simulation Fig. 5.4, the structure of motion in phase space is the same as in the cased of microseisms, Fig. 5.3: loops of different mean radius with the center of the loops oscillating irregularly along a well defined path. Its corresponding power spectrum is presented in Fig. 5.5 and is composed of a main peak at 0.330 Hz (corresponding to $\alpha_0 = -4.0$) and another located at 0.198 Hz. The peak located at 0.330 Hz corresponds to the resonant frequency of the potential, whereas the peak located at 0.198 Hz corresponds (slightly shifted) to the external force f_2.

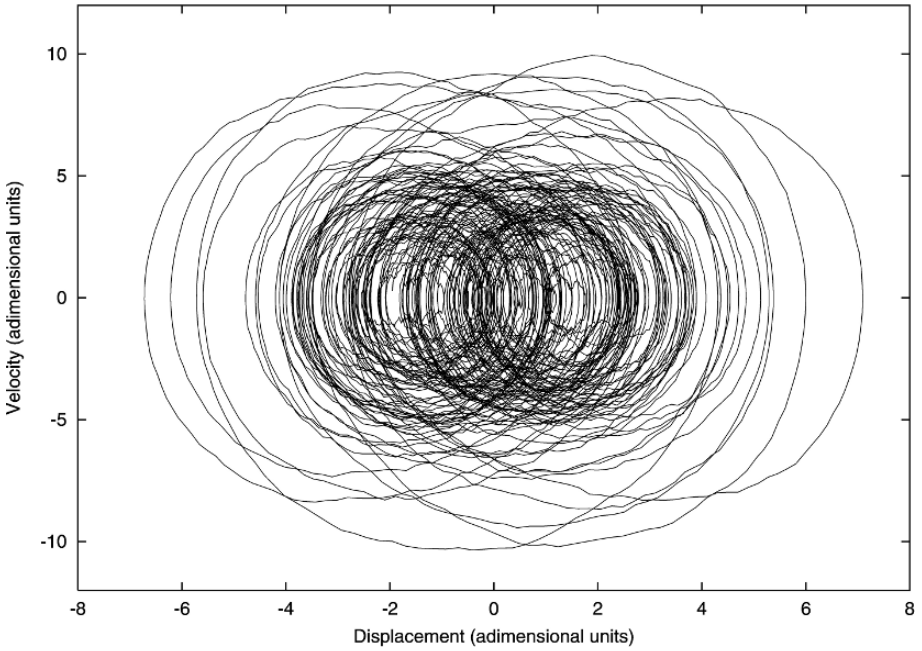


Fig. 5.4. Motion in phase space of a time series generated by the model eq. (2).

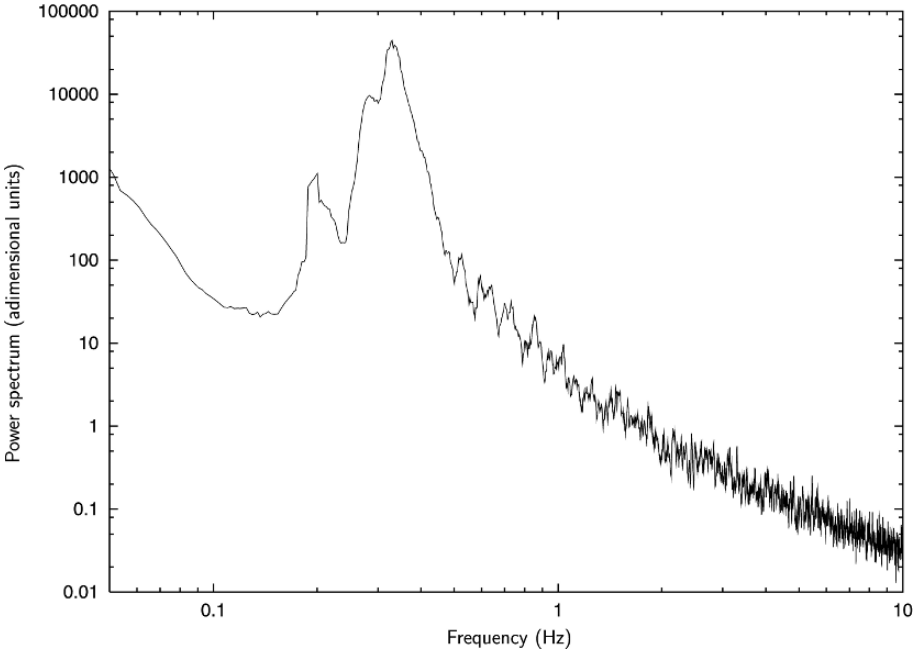


Fig. 5.5. Power spectrum of the time series generated in Fig. 5.4.

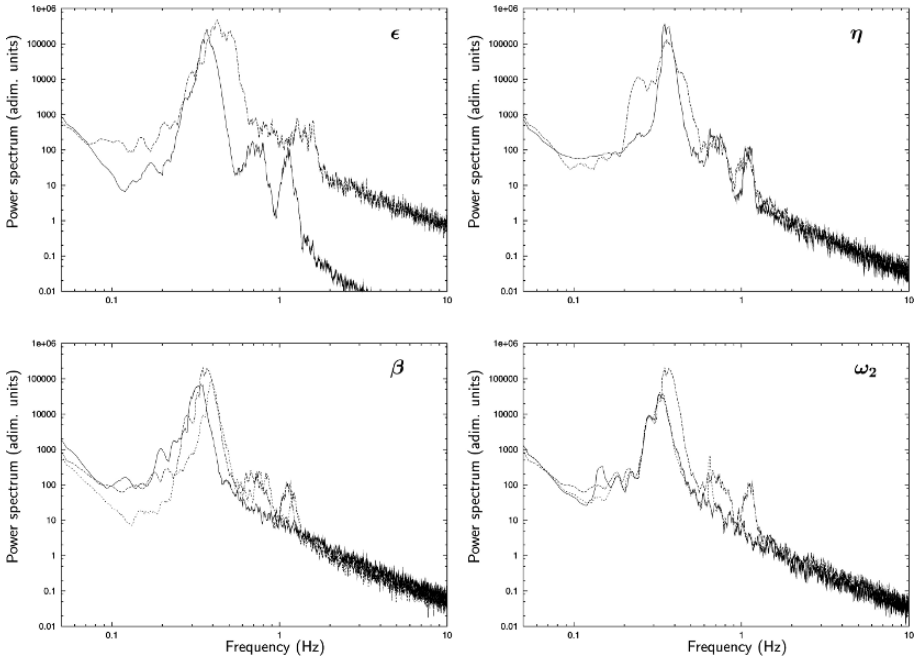


Fig. 5.6. Influence of the variation of the model parameters (amplitude of the white noise, amplitude of the parameter resonance, coefficient of nonlinearity and frequency of the external force) on the power spectrum. It can clearly been observed the variation of the broadness and central frequency of the main peak, as well as growing of sub- and super-harmonics.

Figure 5.6 displays the variations of the power spectrum due to variations of the model parameters. The reference resonant response of the medium has been computed for the following parameter values: $\alpha_0 = 0$, $\gamma_1 = \gamma_2 = \eta = 0$, $\varepsilon = 1.0$, $\delta = 0.01$ and $\beta = 0.05$; the resonant frequency is located at $f_r = 0.34$ Hz. This reference resonant response is then perturbed giving different values to the rest of parameters. By introducing the values $f_1 = 0.05$, $\gamma_1 = 7.5$, $f_2 = 0.35$, $\gamma_2 = 1.0$ and $\eta_2 = 0.03$, the shape and frequency of the resonant peak are preserved whereas new high frequency peaks have been generated. The power spectrum corresponding to these values have been taken as the reference in Fig. 5.6, drawn as a solid line. In this figure at the top in the left we display the change in the power spectrum due to higher amplitude of the white noise (ε): the new spectrum is clearly broadened and the main peak has shifted to higher frequencies. In the top at the right, for a higher value of the amplitude of the parametric resonance (η) the spectrum broadens, although the frequency of the main peak is preserved. At the bottom left we can see that small variations of the nonlinear coefficient (β) induce important effects on the wideness of the main spectral peak, a shifting of its frequency as well as the generation of sub-harmonics and super-harmonics. Finally, at the bottom right we present the influence of the

frequency of the harmonic force f_2 when close to the proper frequency of the potential well, *i.e.,* close to resonance: the main peak widens considerably, its frequency is shifted, and, as in the case of the nonlinear coefficient, sub- and super-harmonics appear.

It has also been observed, through numerical simulations with the simple Duffing oscillator ($\gamma_1 = \varepsilon = \eta = 0$ in eq. (2)) that the time series i) becomes non-stationary when approaching the resonance and ii) becomes stochastic for $\varepsilon \neq 0$. These observations could explain the non-stationarity and stochasticity of the observed microseism time series. The stochasticity can be easily understood because of the random component of the high frequency local meteorological conditions. However, the origin of non-stationarity, although simple in simulations, is not so trivial for real data. The main question is whether there exists a resonant response of the medium, as in the case on a nonlinear oscillator. According to the classical point of view, the response would be negative because the spectral peaks correspond to source terms. Let's see that this may not be the case.

3 Earth's Background Fluctuations

(Vila, Macià, Kumar, Ortiz, Moreno, and Correig 2006) carried out a study on microseism activity at its lower level of excitation, for particular sites where changes in site response are expected. In an attempt to study the stability of the spectral peaks for other areas, a set of 34 world wide broadband seismic stations were individually analyzed for a minimum interval of time of one year in successive windows of 30 minutes (Vila, Macià, Sleeman, and Correig 2007). The power spectrum of each window was corrected by instrumental response. The series of spectra were sequentially compared and the frequency bin of lowest amplitude was selected, thus defining a "pseudo-spectrum": it is not a real spectrum because it consists on the envelope of the minimum amplitude of each bin, a lower bound for each frequency. A real minimum spectrum is also computed by selecting the spectrum of minimum energy in the frequency range of microseism activity. From a numerical point of view, both spectra are hardy distinguishable. While the minimum pseudo-spectrum is useful to determine the medium structural response, the minimum spectrum accounts for the dynamic characteristics of the resonant response.

By comparing the pseudo-spectra of the distinct stations (see Fig. 5.7), it was found that for continental stations located at the northern hemisphere (black lines), all spectra coalesce to a common shape and amplitude, being characterized by a primary peak located at ~0.7 Hz and a secondary peak located at ~0.21 Hz. Continental seismic stations located at the southern hemisphere (grey lines) share the same shape as those of the northern one but with a higher amplitude. (Vila et al. 2007)

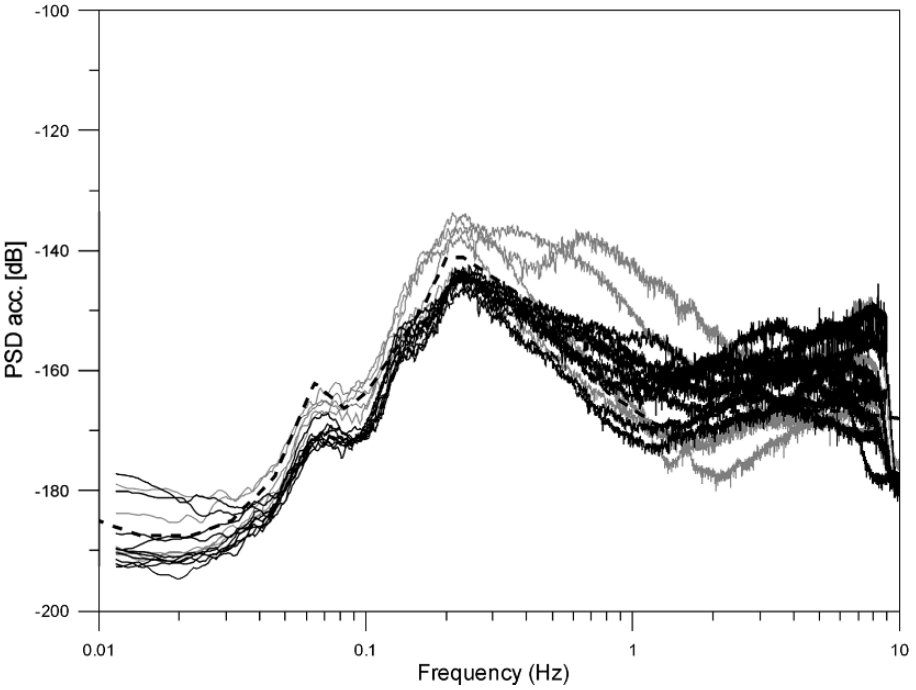


Fig. 5.7. Minimum pseudo-spectra of the time series recorded at several continental seismic stations located at the northern (black lines) and southern (grey lines) hemispheres. In the frequency range 0.04 Hz – 0.3 Hz, the spectra of all northern hemisphere continental stations coalesce, whereas for the stations located at the southern hemisphere the amplitudes of the spectra fluctuates, although preserving the frequency peaks. For comparison, the reference Low Noise Model by (Peterson 1993) is also plotted (black dashed line).

found that the non continental stations suffer strong fluctuations but sharing the same spectral peaks. Because it was selected the spectra of minimum amplitude, the influence of the sources should be highly diminished, and so the minimum spectra can be considered as representative of the medium response, a global property. As the energy is proportional to the square of the amplitude, the minimum power spectrum can be understood as the spectrum corresponding to a wave field at its lower energy level for which all transients (that propagate at the ballistic regime) have vanished after a process of multiple scattering to enter in a diffusive regime (characterized by a diffusive decay with time) and further to the equipartioning regime (for which the flux of energy falls to zero, but the average energy is still above the background level), and for which the wavefield has lost any preferable propagation direction (Hennino, Tregoures, Shapiro, Margerin, Campillo, van Tiggelen, and Weaver 2001).

In order to derive a phenomenological model to simulate the recorded wave field at its minimum energy, that is, the earth's background fluctuations, we hypothesize i) the observations correspond to the diffuse/equipartitioned regime, ii) the energy needed for the excitation of the wave field is that remaining after all transients have vanished, that is, the coda of body- and surface-waves generated from earthquakes and oceanic storms, and iii) generalizing previous observations of coda waves in the equipartitioned regime.

It can be shown (Milotti 1995; Sinha 1996) that $1/f^{\xi}$ noise, actually accepted as the only source term, can be interpreted as a superposition of transients, in good agreement with the (presently generalized) classical coda model of (Aki and Chouet, 1975). Thus, we model the transient spectrum of the energy as a simple exponential-relaxation process.

$$\begin{aligned} N(t) &= N_0 e^{-\lambda t}, \quad t \geq 0 \\ &= 0, \qquad\quad t < 0 \end{aligned} \tag{3}$$

As the energy, in the form of wave-packets, is continuously generated, we consider a summation of exponential process with the inter-event time following a Poisson distribution and assume, for simplicity, initial amplitude N_0 constant for each exponential process; also, a random phase has been added.

Our proposed model follows the model eq. (2) with all force terms substituted by eq. (3) and with the term $V(q)$ as in eq. (2) but for $\alpha(t) = \alpha_0$, an so rewriting $V(q) = V_0(q)$:

$$\frac{d^2q}{dt^2} + \frac{dV_0}{dq} + \delta\frac{dq}{dt} = N_0 \sum_{i=1}^{n} e^{-\lambda(t-t_i)} \tag{4}$$

The sub-index i stands for any one of the n wave-packets.

Figure 5.8 displays the spectrum corresponding to the force term, the right-hand-side of eq. (4) for the following values of the parameters: $\lambda = 0.01$, $N_0 = 5.0$ and for the Poisson distribution $P(t_i - t_{i-1}) = exp[-1.0(t_i - t_{i-1})]$. Clearly a $1/f^{\xi}$ has been obtained, with $\xi = 1.8$.

Taking into account the above source term, the model eq. (4) has been solved for $\alpha_0 = 4.0$, $\delta = 0.01$, and the following values of the non-linear parameter: $\beta = 0.0$, 0.05 and 1.0. The corresponding results are displayed in Fig. 5.9. For $\beta = 0.0$ the main peak is located *ca*. 0.3 Hz, and progressively shifts to higher frequencies as β increases. One sub-harmonic located at *ca*. 0.2 Hz is always present, becoming stronger as β increases, probably due to a competition between the two non-linear

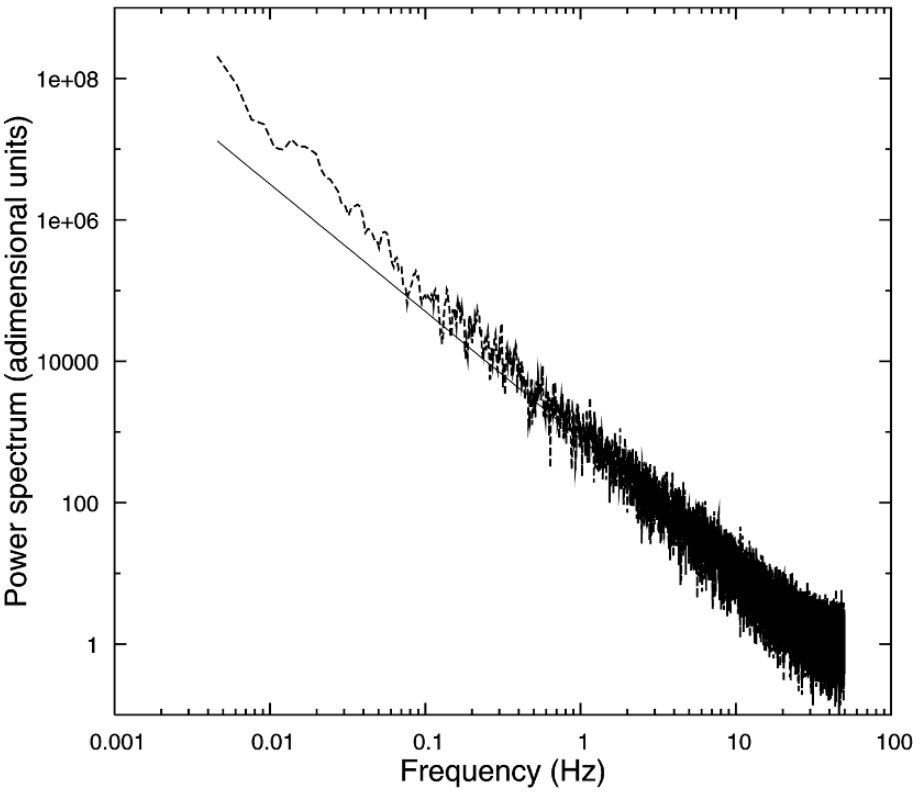


Fig. 5.8. Example of the source term $1/f^{\xi}$ noise of model eq. (4).

terms β and $1/f^{\xi}$. Note that although the main peak strongly depends on β, the sub-harmonics are invariant. The insert of Fig. 5.9 displays our selected model, corresponding to a low nonlinearity. The motion in phase space closely follows that of Fig. 5.4, thus displaying the non-linear observed behavior.

4 Discussion and Conclusions

In the absence of earthquakes, the earth is in a state of continuous unrest, called hum for low frequencies (from 0.002 Hz to 0.01 Hz), microseism activity (from 0.01 Hz to 1 Hz) and seismic local noise (from 1 Hz). At frequencies less than those of the free oscillations of the Earth (less than 1 hour), this unrest accounts for variations of the polar motion. In the present study we have centered in the microseism activity frequency range, strictly speaking for the interval 0.04 Hz to 0.3 Hz to avoid the influence of oceanic infragravity waves as well as the regional and local medium resonances. For this frequency range the major source of unrest lies in the influence of atmospheric storms over oceanic areas.

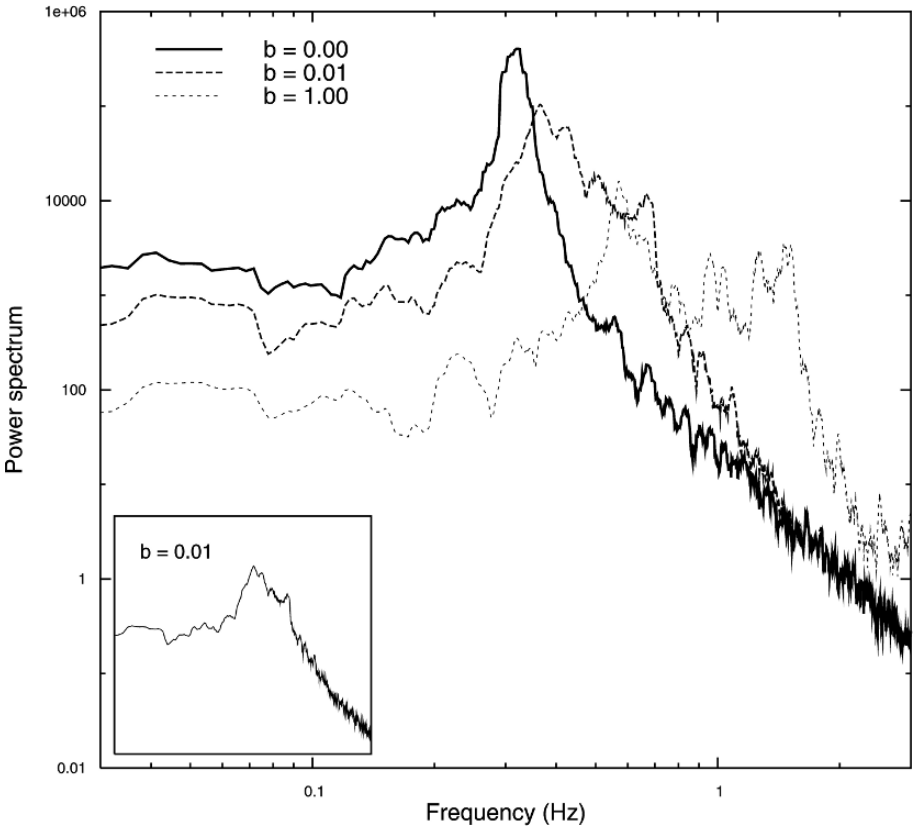


Fig. 5.9. Power spectrum generated by model eq. (4) and its evolution for different values of the non-linear coefficient. Although the shift of the main peak to higher frequencies as a function of the nonlinear coefficient, in all cases the subharmonics of the minimum spectra are preserved. The insert displays our selected model, corresponding to a low nonlinearity.

We have seen that the power spectrum of microseism activity is quite similar in stormy weather conditions and in quite days, consisting in a primary peak located at about 0.07 Hz and a secondary peak located at about 0.2 Hz. However, in stormy conditions the amplitude and frequency of both peaks may suffer severe variations. Two phenomenological models are proposed. The first one takes into account the well developed microseism activity, generated by atmospheric stormy conditions, and is represented by equation (eq. (2)) with external harmonic forces and noise. This model accounts for the stochasticity of the observed time series (due to the presence of white noise), its nonlinearity (the model needs to be nonlinear to reproduce observations) and the non-stationarity (that can be interpreted as due to the competition of the external forces with the resonant medium response). The resonant medium response can be computed by making all external forces vanish except for a small amplitude white noise. We interpret this finding in the following way: when

the external forces are at its minimum level, the power spectrum reflects the medium response, and when they are actively present (the case of active meteorological storms) the observed power spectrum is a manifestation of a competition of the medium response with the external forces. This interpretation is in contradiction with the generally accepted point of view that the spectral peaks are a manifestation of the external forces.

In order to get more insight on the previous suggestion, a study has been carried out on world wide microseism activity at its lower level, that is, in the absence of transients, and with the external forces practically vanished. Surprisingly, the spectra of all continental stations practically coalesce to a common curve for the whole interval of time analyzed (six years in our study for the CADI station, for example), a strong evidence in favor of that the spectral peaks are independent of the external forces or, in other words, that these spectral peaks are representative of the medium response. Moreover, the fact that observations at very different tectonic settings present the same characteristics (the spectral peaks located always at the same frequencies) strongly suggests that the characteristics of the power spectrum at the minimum energy level is a global property, independent of the local structures. For the case of stations located at oceanic islands or continental margins, the shape of the spectra is the same (that is, the primary and the secondary peaks are located at the same sites), although the amplitudes fluctuate. These variations can be explained in the following way: the southern hemisphere is mainly composed of oceans with permanent and strong oceanic currents, hence with a non-negligible external force. Extrapolating the whole set of results, we can safely conjecture that we are indeed in the presence of a global medium response, which fluctuations due to small variations of the steady state energy budget, we have termed Earth's fluctuations.

A phenomenological model has been devised to account for them (see eq. (4)), in which we have hypothesized that the energy needed to excite the Earth fluctuations can be characterized as $1/f^{\xi}$ noise. The justification of this election is as follows, based on the energy liberated by the dominant excitation mechanisms: *i)* at very low frequencies (much lower than the microseismic interval) the combination of atmospheric and oceanic processes, atmospheric circulation and ocean bottom pressure fluctuations, are able to excite the Chandler wobble (Gross, Fukumori and Menemenlis 2003) with a characteristic period of about 14 months, *ii)* at lower periods, between 2 mHz and 7 mHz, still lower than, microseismic activity, severe storms located mainly at the Pacific Ocean are able to excite Rayleigh waves of large periods (Rhie and Romanovicz 2004), *iii)* at the frequency interval of microseismic activity, 0.04 through 0.3 Hz, the excitation functions, less energetic, are the ocean-bottom pressure fluctuations originated by local-regional atmospheric storms over the oceans (Longet-Higgins model) and the surf mechanism, as well as the remaining codas of body and surface waves generated by earthquakes and, *iv)* at higher frequencies, much less energetic, the excitation mechanism is provided by local meteorological conditions. Hence, we find it reasonable, as a first approximation, to model the source as $1/f^{\xi}$ noise. Note that in the presence of atmospheric activity, our model is represented by equation (1), in which Longet-Higgins source term is included as an external force. It is worth to say that we could have used directly our

model (eq. (2)) with all source terms vanishing except the white noise term, obtaining similar numerical results. However, it is hard to accept white noise as a source of the equilibrium response. For this reason we propose as a source the $1/f^{\varepsilon}$ noise model, able to account for all possible sources of energy and able to account for observations. At this point it is worth to note the growing interest of the influence of the dynamics of the atmosphere upon the dynamics of the lithosphere, as for example the study of (Lowry 2006) on the influence of climate load stress upon fault slips.

From a generic point of view we can compare the recorded equilibrium fluctuations wave field to black-body radiation. Black-body radiation refers to a system which absorbs all radiation incidents upon it and re-radiates energy which is a characteristic of this radiation system only, not depending upon the type of radiation which is incident upon it. The energy absorbed by the earth is that of the atmosphere, the oceans and earthquakes. When observing the minimum energy spectrum, the input of energy can be considered as quasi-stationary and isotopic. In the actual case, instead of temperature we can talk about stress. As the Earth is an open system, the incoming energy is re-radiated via a scattering process, giving rise to the background fluctuations with a well-defined reference spectrum that corresponds to the medium response, that can be associated, for a fixed point, to a chaotic oscillator in the presence of noise, giving rise to the resonance of the medium in the frequency interval of the microseism activity at its lower level. Obviously we do not deny the existence of external forces such as those of Longet-Higgins (originated by oceanic storms) and the surf mechanism. What we claim is that these harmonic forces, more or less close to the medium resonance, are not represented by the two spectral peaks but compete with them giving rise to the so often observed strong fluctuations of the medium spectral peaks.

In short, we can say that the spectral peaks of the Earth's equilibrium fluctuations are a manifestation of the medium resonant response, being the following the evidences of this interpretation: *i)* the global stability of the minimum spectra, *ii)* the motion of the system in phase space and *iii)* the frequency ratio 3:1 of the spectral peaks revealed from the whole set of minimum spectra, instead of the frequency ratio 2:1 predicted by the previous models. When external forces are present (atmospheric and oceanic), that is, in the case of large microseism activity, a competition between these forces and the medium resonances develop, resulting in severe fluctuations of the spectral peaks in both amplitude and main frequencies.

To conclude, we can say that the monitoring of the minimum energy spectrum may have potential applications in getting insight on geophysical dynamical process such as the variation of the stress field in volcanic processes (Garcia, Vila, Ortiz, Macià, Sleeman, Marrero, Sánchez, Tárraga and Correig 2006; Vila *et al*. 2006) or (from gravity observations) hydrothermal processes in geysers (Tikku, McAdoo, Schenewerk, and Willoughby 2006).

Acknowledgments

The facilities of the IRIS Data Management System, and specifically the IRIS Data Management Center, were used for access to waveform and metadata required in this

study. The IRIS DMS is funded through the National Science Foundation and specifically the GEO Directorate through the Instrumentation and Facilities Program of the National Science Foundation under Cooperative Agreement EAR-0004370. This research has been supported in part by the Spanish Ministry of Education and Science under Grants CGL2003-21643-E-TEGETEIDE and HG2004-0022. Contribution n. 305 of the Departament d'Astronomia i Meteorologia of the Universitat de Barcelona.

References

Aki, K. and Chouet, B. (1975) Origin of coda waves: source, attenuation, and scattering effects, J. Geophys. Res., 80, 3322-3342.

Bath, M. (1973) *Introduction to seismology*, Birkhauser Verlag, Basel.

Bendat, J.S. and Piersol, A.G. (1986). *Random data analysis and measurement procedures*, John Wiley & Sons.

Correig, A.M. and Urquizú, M. (2002) Some dynamical characteristics of microseism time series, Geophys. J. Int. 149, 589-598.

Correig, A.M., Urquizú, M., Macià, R. and Vila, J. (2006) $1/f^{\alpha}$ noise as a source of the Earth's fluctuations, Europhys. Lett., 74, 581-587.

García, A., Vila, J., Ortiz, R., Macià, R., Sleeman, R., Marrero, J.M. Sánchez, N., Tárraga, M. and Correig, A.M. (2006) Monitoring the reawakening of the Canary Islands' Teide Volcano, *EOS*, 87 (6), 61-72.

Gross, R.S., Fukumori, I. & Menemenlis, D. (2003) Atmospheric and oceanic excitation of the Earth's wobbles during 1980-2000, J. Geophys. Res., 108, ETG 6-1-ETG 6-16

Hasselmann, K. (1963) A statistical analysis of the generation of microseisms, Rev. Geophys., 1, 177-210.

Hennino, R., Tregoures, N., Shapiro, N.M. Margerin, L. Campilli, M., van Tiggelen, B.A. and Weaver, R.L. (2001) Observation of equipartition of seismic waves, Phys. Rev. Lett. 86, 3447-34350.

Lachet, C. and Bard, P. Y. (1994) Numerical and theoretical investigations on the possibilities and limitations of Nakamura's technique, J. Phys. Earth, 42, 377-379.

Longet-Higgins, M.S. (1950) A theory for the generation of microseisms, Phil. Trans. R. Soc., Ser A, 243, 1-35.

Lowry, A.R. (2006) Resonant slow fault slip in subduction zones forced by climatic load stress, Nature, 442|17, 802-805, doi:10.1038/nature05055.

Milotti, E. (1995) Linear processes that produce 1/*f* or flicker noise, Phys. Rev. E, 51, 3087-3103.

Peterson, J. (1993) Observations and modelling of seismic background noise, U.S. Department of Interior Geological Survey, Open file report 93-322, 95 pp. Albuquerque, New Mexico.

Press, F. and Ewing, M. (1948) A theory of microseisms with geologic applications, Trans. Am. Geophys. Union, 29, 163-174.

Rial, J. (1989) Seismic wave resonances in 3-D sedimentary basins. Geophys. J. Int., 99, 81-90.

Rhie, J. and Romanovicz, B. (2004) Excitation of the Earth's continuous free oscillations by atmosphere-ocean-seafloor coupling, Nature 41, 552-556.

Ryabov, V.R., Correig, A.M., Urquizú, M. and Zaikin, A.A., 2003. Microseism oscillations: from deterministic to noise driven models, *Chaos, Solitons and Fractals*, 16, 195-210.

Sinha, S. (1996) Transient 1/*f* noise, Phys. Rev. E, 53, 4509-4513.

Tikku, A.A., McAdoo, D.C. Schenewerk, M.S. and Willoughby, E.C.(2006) Temporal fluctuations of microseismic noise in Yellowstone's Upper Geiser Basin from a continuous gravity observation, Geophys. Res. Lett. 33, L11306, doi : 10.1029/2006GL026113.

Vila, J. (1998) The broad band seismic station CAD (Tunel del Cadi, eastern Pyrenees): site characteristics and background noise, Bull. Seism. Soc. Am., 88, 297-303.

Vila, J. and Macià, R. (2002) The broad band seismic station CADI (Tunel del Cadi, eastern Pyrenees), part II: long period variations of background noise, Bull. Seism. Soc. Am., 92, 3329-3334.

Vila, J., Macià, R., Kumar, D., Ortiz, R. Moreno, H. and Correig, A.M. (2006) Analysis of the unrest of active volcanoes using variations of the base level noise seismic spectrum, J. Volc. Geotherm. Res., 153, 11-20.

Vila, J., Macià, R., Sleeman, R. and Correig, A.M. (2007) World wide base level noise seismic spectrum, J. Geophys. Res. (submitted).

Webb, S.C. (1998) Broadband seismology and noise under the ocean, Rev. Geophys, 36, 105-142.

Wiechert, E. (1904) Verhandlurgen der zweiten Internationaler Seismologischen Konferenz, Gerl. Beit. Geophys. Ergonzvngsband, 20.

6 An Exponential Langevin-type Model for Rainfall Exhibiting Spatial and Temporal Scaling

Victor B. Sapozhnikov[1] and Efi Foufoula-Georgiou[1]

[1] St. Anthony Falls Laboratory, National Center for Earth Surface Dynamics, Department of Civil Engineering, University of Minnesota, Minneapolis

Abstract. Models for simulating the space-time structure of rainfall over a range of scales remain a valuable tool for hydrologic and geomorphologic studies. A desired attribute of these models is the preservation of statistical properties of observed precipitation fields, among which are well documented spatial and temporal scaling. In this paper we propose a class of stochastic space-time rainfall models based on exponentiation of evolving nonlinear Langevin-type surfaces. The proposed models have the potential of reproducing (a) the spatial and temporal scaling structure of observed rainfall fields, (b) the spatial and temporal intermittency in rainfall over a range of scales, and (c) possible nonstationarity in space and time.

1 Introduction

Precipitation varies simultaneously over space and time. Although several studies have concentrated on the scaling structure of spatial rainfall fields and separately on the scaling structure of temporal rainfall (See Lovejoy and Schertzer 1990; Gupta and Waymire 1993; Kumar and Foufoula-Georgiou 1993, Carsteanu and Foufoula-Georgiou 1996; among others; see also the review papers of Foufoula-Georgiou and Krajewski 1995; Foufoula-Georgiou 1998), it is obvious that space and time are not independent of each other and should be considered simultaneously. Towards this direction, the studies of Over and Gupta (1996); Marsan, Schertzer, and Lovejoy (1996); and Venugopal, Foufoula-Georgiou, and Sapozhnikov (1999) are worth noting. Over and Gupta (1996) proposed a space-time model based on a multiplicative cascade in space and a stochastic evolution in time. At each time step, the spatial field is generated by using cascade generator weights that are obtained via a Markov process from the weights at the previous time step. Marsan et al. (1996) proposed a space-time multifractal model which uses a three-dimensional anisotropic multiplicative cascade. They used a continuous cascade generated from a white Levy noise fractionally integrated at the smallest scale present. Venugopal et al. (1999) reported evidence of dynamic scaling in normalized rainfall fluctuations and proposed a space-time stochastic downscaling model which can also be used for simulation purposes. An alternative model was recently proposed by Kundu and Bell (2006) based on white-noise-driven fractional diffusion and it was shown to exhibit dynamic scaling.

In this paper, we explore a different class of models for space-time rainfall based on continuum growth equations and specifically the nonequilibrium models described by the Langevin-type equation introduced by Kardar, Parisi and Zhang (1986). These models have been extensively studied in recent years, because they

exhibit interesting scaling properties, are simple and still are able of capturing important features of physical processes such as depositional surface growth (e.g., see Edwards and Wilkinson 1982; Meakin, Ramanlal, Sander, and Ball 1986; Family 1986; Lam and Family 1991), fluid displacement in porous media (e.g., Rubio, Edwards, Dougherty, and Gollub 1989; Kahanda Zou, Farrel, and Wong 1992; He, Kahanda, and Wong 1992), the growth of bacterial colonies (e.g., Vicsek, Cserzo, and Horváth 1990), diffusion-driven evolution of interfaces (Sapozhnikov and Goldiner 1991, 1992), turbulence (e.g, Kardar et al. 1986; Schwartz and Edwards 1992; see also references in Bohr, Jensen, Paladin, and Vulpiano 1998), landscape evolution (e.g., Czirok, Somfai, and Vicsek 1993; Sornette and Zhang 1993; Stomfai and Sander 1997; Banavar, Colaiori, Flammini, Saritan, and Rinaldo 2001; see also Passalacqua, Porté-Agel, Foufoula-Georgiou, and Paola 2006), and others. In this article we use Langevin-type surfaces to build a space-time model for rainfall and demonstrate that it reproduces observed space-time scaling properties.

In the next section, a brief review of the surface growth models and especially the KPZ (Kardar et al. 1986) model and its space and time scaling characteristics is given. In section 3, an adaptation of the KPZ model for space-time rainfall simulation is proposed and shown to reproduce certain empirical scaling characteristics found in rainfall. Finally conclusions and directions for further research are discussed in section 4.

2 Surface Growth Models

The geometry and evolution of surfaces formed by stochastic processes causing roughening and smoothing, e.g., deposition, dendritic growth and diffusion-driven evolution of interfaces, has been studied theoretically and by computer simulation (see, e.g. Edwards and Wilkinson 1982; Kardar et al. 1986; Meakin et al. 1986; Family 1986; Lam and Family 1991; Sapozhnikov and Goldiner 1992; and Barabasi and Stanley 1995). These surfaces have been shown to exhibit double scaling, both in space and in time. Let us denote by $h(\mathbf{x},t)$ the height of the surface at location $\mathbf{x}$ and at time t, by $\overline{h(\mathbf{x},t)}$ the average height of the surface over the whole system of size L, and by $\sigma(L,t)$ the surface width, which characterizes the roughness of the surface, defined as the root mean square (RMS) fluctuation in height,

$$\sigma(L,t) \equiv < [h(\mathbf{x},t) - \overline{h(\mathbf{x},t)}]^2 >_x \tag{1}$$

where $<>_x$ denotes spatial (over $\mathbf{x}$) average over the size of the system L. For many growth surfaces it has been found that initially the width increases as a power of time

$$\sigma(L,t) \sim t^{\beta}, \qquad t \ll t_{\times} \tag{2}$$

where β is called the growth exponent and characterizes the time-dependent dynamics of the roughening process. The power law increase of width continues up

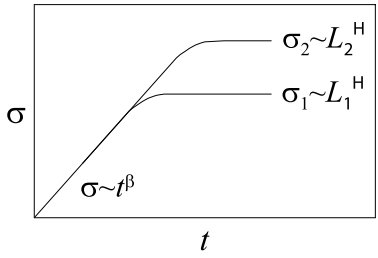


Fig. 6.1. Schematic of the log-log dependence of the $h(x,y,t)$ surface width σ on time, for two different linear sizes of the lattice L. Each of the two lines saturates at $\sigma \sim L^H$.

to a time $t_\times$ (called the cross-over time) at which the width reaches a saturation value σ_{sat} which depends on the size of the system. As the size L of the system increases, so does σ_{sat} and the dependence follows a power law

$$\sigma_{sat}(L) \sim L^H, \qquad t \gg t_\times \tag{3}$$

The exponent H is called the roughness exponent and characterizes the roughness of the saturated surface.
The cross-over time $t_\times$ at which the surface crosses from the behavior of (2) to that of (3) depends on the system size

$$t_\times \sim L^z \tag{4}$$

where z is called the dynamic exponent and can be shown to be

$$z = H/\beta \tag{5}$$

The above equations can be combined into one double-scaling relation (Family and Vicsek 1985; or Barabasi and Stanley 1995)

$$\sigma(L,t) = L^H f(t/L^z) \tag{6}$$

where $f(\xi) \sim \xi^\beta$ as $\xi \to 0$ (giving $\sigma \sim t^\beta$) and $f(\xi) \to const$ as $\xi \to \infty$ (giving $\sigma \sim L^H$) (see Figure 6.1).

It has also been shown that the H exponent is the Hurst exponent of the $h(\mathbf{x},t)$ surface at any time t_0 after σ stabilizes, and thus describes the geometry of this surface at any given instance of time t_0. Furthermore, as discussed later, β is equal to the Hurst exponent H_t of the surface evolution at any fixed point $\mathbf{x}_0$, i.e., of $h(\mathbf{x}_0,t)$. Based on the scaling exponents β and H, such surfaces have been found to fall into distinct classes of universality, depending on the character of the processes that give rise to the surface evolution (e.g., see Liu and Plischke 1988; Krug 1987; Barabasi and Stanley 1995).

To connect the shape and evolution of growing surfaces to the forming processes, Edwards and Wilkinson (1982) suggested a linear equation of Langevin type and showed that it yields the spatial and temporal scaling of the surface. This work was later extended by Kardar et al. (1986). They proposed the following nonlinear equation of Langevin type for surface growth (often referred to as the KPZ equation)

$$\frac{\partial h(\mathbf{x},t)}{\partial t} = \nu\nabla^2 h(\mathbf{x},t) + (\lambda/2)(\nabla h(\mathbf{x},t))^2 + \eta(\mathbf{x},t) \tag{7}$$

where $h(\mathbf{x},t)$ measures the height of the surface relative to the average height and ν and λ are model parameters. The first term in the right-hand side describes the surface smoothing, the second (nonlinear) term is caused by slope dependence of the surface growth rate, and the third term is white noise describing roughening of the surface by a random process. In case $\lambda = 0$ this relation is reduced to the linear equation of Edwards and Wilkinson (1982), which gives $H = 1/2$ and $\beta = 1/4$ for 1-dimensional surfaces $h(x,t)$. Kardar et al. (1986) solved equation (7) analytically for 1-dimensional surfaces $h(x,t)$ and showed that if $\lambda \neq 0$ the exponents in (6) are $H = 1/2$, and $\beta = 1/3$.

To our knowledge equation (7) has not been solved analytically for two-dimensional surfaces $h(x,y,t)$. However, theoretical consideration of equation (7) (Halpin-Healy 1989) and computer simulation (Meakin et al. 1986; Hirsh and Wolf 1986; Liu and Plischke 1988; see also summary Table in Barabasi and Stanley 1995 p.83) suggest that in this case H is estimated to be in the range of 0.36 to 0.385, β in the range of 0.23 to 0.24 and consequently z is approximately 1.6 (Note that these values agree with the values obtained from the equations $H = 2/(d+3)$ and $\beta = 1/(d+2)$ proposed by Kim and Kosterlitz 1989). In subsequent works of Medina, Hwa, Kardar, and Zhang (1989) and Lam and Family (1991) correlated noise was introduced instead of white noise in equation (7). In this case, the scaling exponents were shown to depend on the noise correlation structure. Moreover, Meakin et al. (1986), Krug (1987), and Medina et al. (1989) provided arguments that the nonlinear $(\lambda \neq 0)$ version of Langevin equation (7) yields the scaling exponents H and β connected by the simple relation

$$H + H/\beta = 2 \tag{8}$$

for both correlated and uncorrelated noise and in space of any dimension. Equation (8) provides a relation between the two scaling exponents and has been confirmed by the renormalization group calculation (see Barabasi and Stanley, Chapter 7 1995).

The linear version of the Langevin-type model in two dimensions exhibits no spatial scaling. As shown by Edwards and Wilkinson (1982), the surface produced exhibits logarithmic divergence. Thus this type of model would not be appropriate for rainfall modeling. Therefore we have adopted the nonlinear Langevin-type equation (7), which is also referred to as the KPZ equation, as the building block for the development of a space-time rainfall model.

3 Rainfall Model

Several models exist for simulating the nonlinear Langevin-type equation (7), such as the ballistic deposition model (Family and Vicsek 1985), the Eden growth model (Hirsh and Wolf 1986), or the single-step model of surface growth (Meakin et al. 1986; Liu and Plischke 1988) [The reader is referred to Barabasi and Stanley 1995 for details and other models]. Here, the single step model was chosen because of its simplicity and computational efficiency.

In the single-step model, first a "chessboard" initial surface is produced on a square lattice as

$$h(x, y, 0) = 0.5[1 + (-1)^{x+y}] \tag{9}$$

so the heights of neighboring columns differ by ± 1. Then sites of the lattice are chosen randomly, with equal probability, and if the chosen site is local minimum its height is increased by 2. One can notice that under this deposition procedure the heights of neighboring columns at any time can differ only by ± 1. Meakin et al. (1986) found that this model shows double scaling, with exponents $H = 0.36$, $z = 1.64$ (and thus $\beta = H/z = 0.22$). Liu and Plischke (1988) obtained similar results ($H = 0.375$, $z = 1.625$, and consequently $\beta = 0.23$).

As explained before, the surface average width σ stops increasing after long enough evolution time (indeed, in equation (6) $f(\xi) \to const$, e.g., σ stops depending on t, as $\xi \to \infty$). Usually the Langevin-type models of surface growth are studied at nonstationary stages, when the average width σ of the interface still grows. To simulate rainfall however, we first let the surface evolve enough such that its average width σ reaches the asymptotic value, and then follow the evolution of the surface.

It is an interesting result (e.g., see chapter 2 of Barabasi and Stanley 1995), verified also by us via computer simulation of surfaces corresponding to the linear and the nonlinear versions of the Langevin equation, that after the surface width σ reaches its limiting value, the β exponent, instead of describing the surface growth as in (6), determines how the height of the surface $h(x_0, y_0, t)$ at any point (x_0, y_0), fluctuates with time, e.g., that β is actually the Hurst exponent H_t of the $h(x_0, y_0, t)$ process. We will make use of this result for modeling space-time rainfall by the evolution of the stationary surfaces produced by the KPZ equation upon saturation.

3.1 Stationary rainfall

In the original single-step model, the surface average height increases linearly with time (it does not increase if one introduces evaporation with the same probability as deposition, but this, as found by Liu and Plischke (1988) turns the model into a linear one, with no spatial scaling in two dimensions). The linear growth of the surface

average height would lead to a permanent growth in rainfall intensity with time, which is unrealistic. Therefore, we modified the model to preserve the average surface height, by shifting the surface back as a whole, after each deposition cycle. One could say instead that we considered the surface evolution in the coordinate system bound to the average height of the $h(x,y,t)$ surface. This approach is also consistent with the Langevin equation (7) where $h(\mathbf{x},t)$ measures the height of the surface relative to its average height. Each cycle included N^2 attempts of random deposition, equal to the number of sites in the lattice. Although the average height of the surface and (after the initial growth) the average width σ remained constant, the surface evolved continuously during the simulation. For example, Fig. 6.2(a) displays two cross-sections of the $h(x,y,t)$ surface along the $x-$direction ($y=y_0$), at two instances of time t_1 and t_2, i.e., it shows $h(x,y_0,t_1)$ and $h(x,y_0,t_2)$. The differences seen between the two cross sections are the result of the surface evolution. Alternatively, if the surface evolution is followed over time at one specific point (x_0,y_0), a temporal sequence $h(x_0,y_0,t)$, such as that shown in Fig. 6.3(a) is obtained.

The $h(x,y,t)$ surface was then used to simulate a rainfall field $I(x,y,t)$ as

$$I(x,y,t)=exp(Ah(x,y,t)+B) \tag{10}$$

Here A accommodates arbitrariness of units for $h(x,y,t)$ and B (which amounts to a pre-exponent) arbitrariness of the $I(x,y,t)$ units. It can be shown that the A coefficient is related to the temporal and the spatial scale of the process (e.g., see Barabasi and Stanley 1995). Indeed, if the $h(x,y,t)$ surface exhibits scaling in space and in time, with the exponents H (same for the x and the y - directions) and H_t, respectively, then

$$Ah(x,y,t)=h(A^{1/H}x,A^{1/H}y,A^{1/H_t}t) \tag{11}$$

Thus multiplying the surface height by A is equivalent to its space and time rescaling. Figures 6.2(b,c) show the temporal evolution of two cross-sections of the simulated rainfall field $I(x,y,t)$, at two different time instances. They have been obtained from exponentiation of the $h(x,y,t)$ cross sections shown in Fig. 6.2a, with two different values of A. Similarly, Figs. 6.3(b,c) show two temporal sequences of

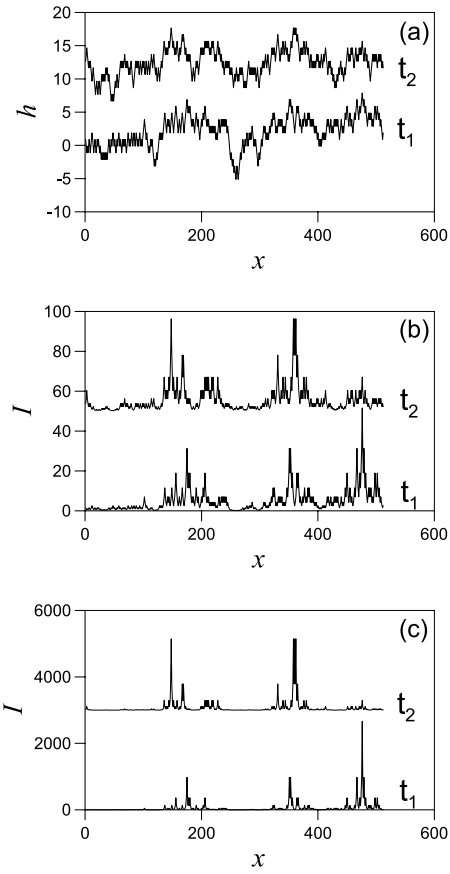


Fig. 6.2. (a) Cross-sections of the evolving surface $h(x, y, t)$ at two different time instances t_1 and t_2, 20 iteration steps apart; (b) Cross sections of the simulated rainfall field $I(x,y,t)$ produced by exponentiation of the $h(x, y, t)$ surfaces whose cross sections are shown in (a) and for $A = 0.5$, $B = 0$; and (c) same as (b) but with $A = 1$, $B = 0$. To avoid overlapping, the cross sections at time $t = t_2$ in (a), (b) and (c) have been shifted upwards by 10, 50 and 3000 units, respectively.

simulated rainfall intensity at a point, for different A and B values. These sequences look similar to observed temporal rainfall sequences shown in Figs. 6.3(d,e) for two Iowa storms in November 1, 1990 (see Georgakakos, Carsteanu, Sturdevant, and Cramer 1994 for description of these data; and also Venugopal and Foufoula-Georgiou 1996 for an extensive wavelet analysis of these storms).

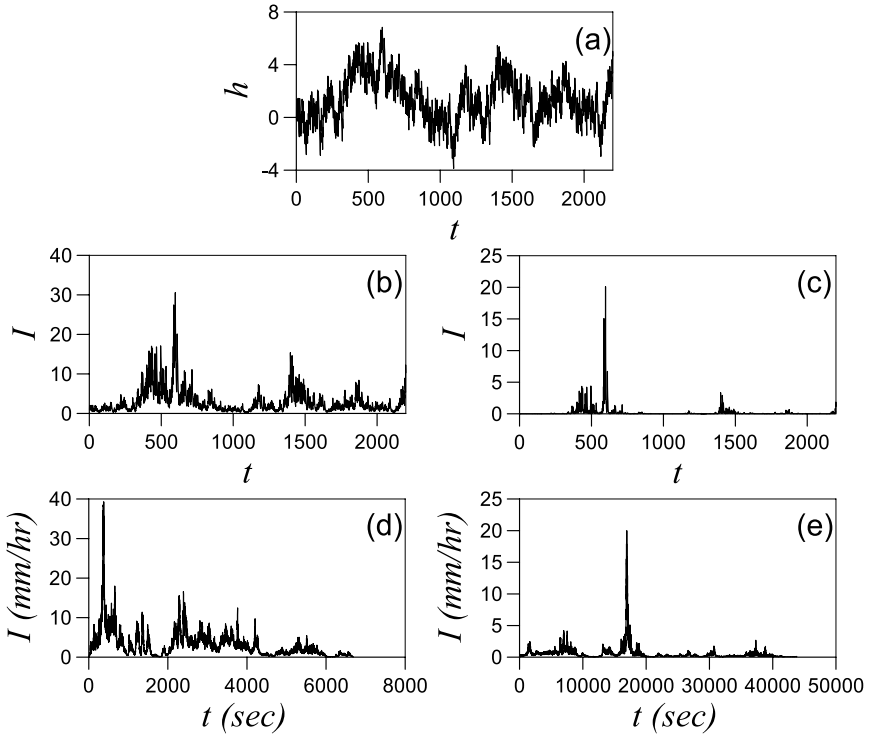


Fig. 6.3. (a) Temporal evolution of the simulated surface $h(x_0, y_0, t)$ at one arbitrary point (x_0, y_0); (b) temporal evolution of the simulated rainfall intensity $I(x_0, y_0, t)$ at point (x_0, y_0) and for $A = 0.5$, $B = 0$; (c) same as (b) but for $A = 1.3$, $B = -6$; (d) and (e) rainfall intensity $I(x_0, y_0, t)$ for two November 1, 1990 storms (storm A and B, respectively) in Iowa City. Notice the similarity between the simulated and the actual rainfalls ((b) resembles (d), and (c) resembles (e)).

The two-dimensional view of the evolving rainfall field, for three different time moments and two different A values, is presented in Fig. 6.4. One can see in Fig. 6.4 that changing the value of A effectively leads to change in the scale of the rainfall field.

As expected, the logarithm of the simulated rainfall field, e.g. the $h(x, y, t)$ surface, showed spatial scaling. Its increments, $\Delta h(x, y, t)$, scaled in space, and gave a Hurst exponent $H \simeq 0.36$ (Figure 6.5). This is close to the spatial Hurst exponent $H = 0.2 - 0.4$ of the normalized fluctuations of natural rainfall $\Delta I(x, y, t)/I(x, y, t) (\simeq \Delta \log(I(x, y, t)) \equiv \Delta h(x, y, t))$ found by Perica and Foufoula-Georgiou (1995) for a large number of midwestern convective storms, providing thus quantitative validation of the proposed model.

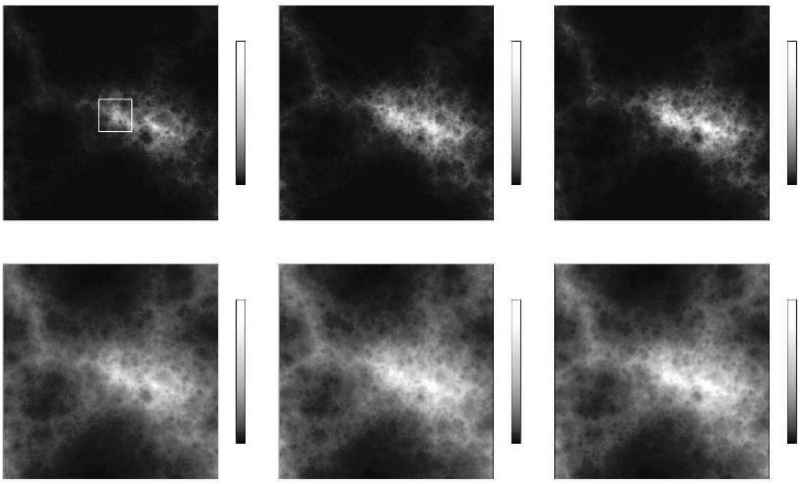


Fig. 6.4. Evolution of the simulated rainfall in two dimensions. The same $h(x, y, t)$ surface (each frame 20 deposition cycles apart from the previous one) is used to produce the rainfall fields in rows (a) and (b) but the A coefficients determining the scale of the rainfall are different: $A = 0.1$ for the upper row, and $A = 0.2$ for the lower row. According to equation (11) the spatial scale in the upper row is $(0.1/0.2)^{1/H} = (0.1/0.2)^{1/0.36} \simeq 7$ times larger than in the lower one. Indeed, one could notice that the rainfall shown in the lower row statistically resembles the boxed smaller part of the upper-row rainfall (box size is 1/7 the size of the whole field). Thus decreasing the A coefficient amounts to zooming into a rainfall field.

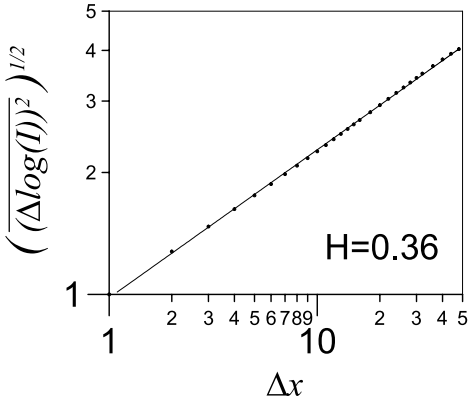


Fig. 6.5. Spatial scaling in the logarithm of the simulated rainfall field, $I(x, y, t)$. The Hurst exponent is $H \simeq 0.36$. It falls within the range $H = 0.2 - 0.4$ found for normalized spatial fluctuations in rainfall (which are approximately the fluctuations in log(rain): $\Delta I/I \simeq \Delta \log(I)$) by Perica and Foufoula-Georgiou (1996).

The logarithm of the simulated rainfall also showed temporal scaling as expected theoretically from the model. The evolving height of any point of the h-surface, $h(x_0, y_0, t)$ behaved as a trail of a fractional Brownian motion, with Hurst exponent H_t = 0.21 (see Figure 6.6a). We compared this result with the behavior of $\log$(rainfall intensity) at a point, for several rainfalls in Iowa. Although not all the logarithmed rainfalls showed good temporal scaling, some of them did. Their Hurst exponent was $H_t = 0.1 - 0.35$, in agreement with the model (which produces $H_t = 0.21$). For example, the temporal scaling in $\log(I(t))$ of the November Iowa storm presented in Fig. 6.3e is displayed in Fig. 6.6b and shows $H_t = 0.29$. The presence of spatial and temporal scaling in the $h(x, y, t)$ surface implies dynamic scaling in the log of simulated rainfall field (the $\log(I(x, y, t))$ surface), with the dynamic scaling exponent $z = H/H_t \simeq 1.7$.

The physical meaning of z is that any feature at the scale L possesses a characteristic lifetime of order $t \sim L^z$, i.e., the feature disappears and/or is replaced by another feature in this time scale. Thus for rainfall, this would mean that a raincell twice as large as another one would dissipate approximately 3.2 (=$2^{1.7}$) times slower, or that it would take a time proportional to $L^{1.7}$ to create or dissipate a rain cell of size L. This result agrees with the dynamic scaling found in Venugopal et al., (1999) for midwestern convective storms (the values of z reported were approximately 1.7). It is noted that in Venugopal et al. (1999) the dynamic scaling exponent was estimated by finding the space-time renormalization of the form $t \sim L^z$ under which the pdfs of $\Delta \log I(L,t) \equiv \log I_L(\tau + t) - \log I_L(\tau)$ remained invariant, where, $I_L(\tau)$ denotes the rainfall intensity at spatial scale L and time instant τ, and t represents the time lag over which the rainfall evolution is measured.

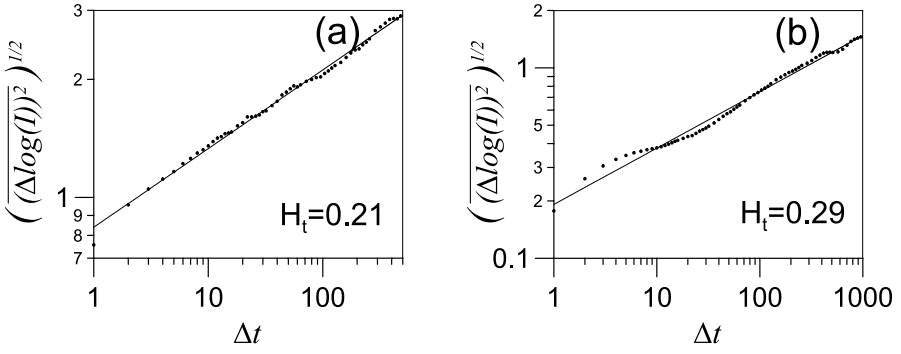


Fig. 6.6. (a) Temporal scaling in the logarithm of the simulated rainfall field $I(x, y, t)$. The Hurst exponent is $H_t \simeq 0.21$. (b) Temporal scaling in the logarithm of the rainfall intensity for the November 1, 1990 B storm in Iowa City. The Hurst exponent is $H_t \simeq 0.29$, close to the exponent produced by the model.

3.2 Nonstationarity and intermittency in rainfall

Whereas the whole rainfall field produced by the proposed model is stationary, its parts are not. Indeed, preserving in time the spatial average of $h(x,y,t)$ over the whole surface does not lead to preserving these averages for an arbitrarily chosen part of the surface. This alone can emulate **nonstationarity** in rainfall. The variations in the average height of a part of the $h(x,y,t)$ surface act as a modulator of the rainfall field intensity $I(x,y,t)$ in that area.

Since at any given instant, different parts of the $h(x,y,t)$ surface have different spatial averages, the rainfall field $I(x,y,t)$ in some of these parts may turn out to be below the sensitivity of the gages (the "dry" areas) and in some other parts above it (the "wet" areas). This creates **spatial intermittency** in the simulated rainfall.

When, as a result of random variation, a point (x_0, y_0) of the evolving $h(x,y,t)$ surface "sinks" deep enough, its exponential mirror, the $I(x,y,t)$ intensity at this point can become very low. We consider time periods over which the rainfall intensity is lower than the sensitivity of a rainfall gage at a given point, as dry periods for this point. When this point "floats up", the next rainfall starts. Thus, the model is capable of reproducing not only a particular rainfall event, but also **temporal intermittency** in rainfall. The temporal intermittency is closely related to the spatial intermittency, both being due to the same features of the generating process.

The model accommodates spatial and temporal intermittency in rainfall by treating dry areas or dry periods as rainfall with very low intensity. Our point is that although some arbitrariness in the sensitivity of the rain gage introduces an element of subjectivity in qualifying a space-time patch as dry or wet, this just reflects the subjectivity existing in real life. Indeed, a sensitive gage could detect rainfall while a less sensitive gage would indicate absence of rain. It should be noted, however, that the exponentiation procedure can easily produce rainfall intensities which are below any reasonably high sensitivity of a gage. Such a period will be dry from the point of view of any observer. Also, it should be mentioned that a space-time patch can be wet or dry depending on its spatial and temporal scale. For example, somebody's hand (which can feel individual drops) put out for a few seconds may not catch any rain drops while a "gage" which is as sensitive as the hand but has a larger size and/or is exposed over a longer time period would signal presence of a low-intensity rainfall.

4 Conclusions

A new class of models for modeling space-time rainfall was introduced as a simple alternative to existing multifractal models. These models are easy to simulate and still have the potential to reproduce important spatial and temporal characteristics of observed rainfall fields over a large range of spatial and temporal scales. These models are based on exponentiation of nonlinear Langevin-type surfaces and represent a rich class within which rainfall of different meteorological environments can be modeled.

In this paper, a simple case was presented in which the standard KPZ model with uncorrelated white noise was used. This model was shown to reproduce well the range of the spatial and temporal scaling exponents empirically documented for midwestern convective storms. Also, by appropriate selection of a model parameter A, the intermittency of space-time rainfall as a function of scale can be reproduced. Since the $h(x,y,z)$ surface is multiplied by the factor A before exponentiation (to produce the rainfall field $I(x,y,t)$) this coefficient affects the spatio-temporal intermittency in the modeled rainfall. In particular, as demonstrated in equation (11), increasing A amounts to increasing spatial and temporal scales. For example, one can notice "dry areas" in Fig. 6.2c obtained using $A = 1.0$, and their absence in Fig. 6.2b obtained from the same $h(x,y,t)$ surface but with $A = 0.5$. Similarly, "dry periods" are present in Fig. 6.3c and absent from Fig. 6.3b. Thus the factor A in the model can be chosen to match the intermittency of a real rainfall field at given spatial and temporal scales.

Although possible advection of the rainfall field imposed by wind has not been presented in this article, the model can easily reproduce it. For that, the rainfall field needs to be shifted in the (x,y) plane with a rate corresponding to the wind velocity.

The model needs further development guided by comparison to real space-time rainfall data, including data on spatial and temporal intermittency. It is noted that, although the parameters A and B can be easily adjusted to reproduce the desired scale and intermittency, the H and H_t exponents describing the space-time structure of the $h(x,y,t)$ surface (and, correspondingly, of the $I(x,y,t)$ rainfall field) cannot be tuned easily. This is because, as most other fractal models, the Langevin-type models exhibit scaling universality, in our case with $H \simeq 0.36$ and $H_t \simeq 0.21$. Although these values pretty much agree with the scaling exponents found in midlatitude convective rain, other scaling exponents (e.g., see Foufoula-Georgiou, 1998 and Foufoula-Georgiou and Venugopal, 2000 for numerically simulated storms and tropical rainfall) can be reproduced using correlated or power-law noise, (e.g., see Medina et al. 1989; Lam and Family 1991; and Zhang 1990), or quenched noise i.e., noise that depends explicitly on h and x but not time t (see chapter 5 of Barabasi and Stanley 1995 and references therein). This issue deserves further investigation.

Acknowledgments

This research was partially supported by the STC program of the National Science Foundation via the National Center for Earth Surface Dynamics (NCED) under agreement EAR-0120914, and by NASA (TRMM and GPM) under agreements NAG5-13639 and NAG5-12909. We thank Venugopal Vuruputur and Sukanta Basu for fruitful discussions. The expert help of David Olsen in typesetting the paper is gratefully acknowledged.

References

Banavar, J., F. Colaiori, A. Flammini, S. Maritan, and A. Rinaldo, Scaling, optimality and landscape evolution, *J. Stat. Physics*, 104, 1-49, 2001

Barabasi A.-L. and H. E. Stanley, *Fractal Concepts in Surface Growth*, Cambridge Univ. Press, Cambridge, 1995

Bohr, T., M.H. Jensen, G. Paladin, and A. Vulpiano, *Dynamical Systems Approach to Turbulence*, Cambridge Univ. Press, United Kingdom, 1998

Carsteanu, A. and E. Foufoula-Georgiou, Assessing dependence among weights in a multiplicative cascade model of temporal rainfall, *J. Geophys. Res.* **(D21)**, 26363, 1996.

Czirok, A., E. Somfai, and T. Vicsek, Experimental evidence for self-affine roughening in a micromodel of a geomorphological evolution, *Phys. Rev. Letters* **(13)**, 2154, 1993.

Edwards, S. and D. Wilkinson, The surface statistics of a granular aggregate, *Proc. Roy. Soc. A* , 17-31, 1982.

Family, F. and T. Vicsek, Scaling of the active zone in the Eden process on percolation networks and the ballistic deposition model, *J. Physics A* , L75-L81, 1985.

Family, F., Scaling of rough surfaces: effect of surface diffusion, *J. Physics A* , L441-L446, 1986.

Foufoula-Georgiou, E. and W. Krajewski, Recent advances in rainfall modeling, estimation, and forecasting, *IUGG*, U.S. National Report 1991-1994 - Contributions in Hydrology, *Rev. of Geophysics*, 1125-1137, 1995.

Foufoula-Georgiou, E. and V. Venugopal, Patterns and organization in precipitation, Ch. 4 in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, R. Grayson and G. Bloschl (eds), Cambridge University Press, United Kingdom, 2000.

Foufoula-Georgiou, E., On scaling theories of space-time rainfall: some recent results and open problems, in: *Stochastic Methods in Hydrology: Rainfall, landforms and floods*, (Eds.) O. E. Barndorff-Nielsen, V. K. Gupta, V. Perez-Abreu, and E.C. Waymire, *Adv. Series Stat. Sci. and Appl. Prob*., 7, 25-72, World Scinetific, 1998.

Georgakakos, K.P., A.-A. Carsteanu, P.L. Sturdevant, and J.A. Cramer, Observation and analysis of midwest rain rates, *J. Appl. Meteorol.*, 1433-1444, 1994.

Gupta, V.K. and E. Waymire, A statistical analysis of mesoscale rainfall as a random cascade, *J. Appl. Meteorology*, 32(2), 251-267, 1993.

Halpin-Healy, T., Diverse manifolfs in random media, *Phys. Rev. Lett.* , 442-445, 1989.

He, S., G. Kahanda, and P-Z. Wong, Roughness of wetting fluid invasion fronts in porous media, *Phys. Rev. Letters*, 69, 3731-3734, 1992

Hirsh, R. and D. Wolf, Anisotropy and scaling of Eden clusters in two and three dimensions, *J. Phys. A* , L251-L256, 1986.

Kahanda, G., X. Zou, R. Farrell, and P-Z. Wong, Columnar growth and kinetic roughening in electrochemical deposition, *Phys. Rev. Letters*, 68, 3741-3744, 1992

Kardar, M., G. Parisi and Y. Zhang, Dynamic scaling of growing interfaces, *Phys. Rev. Lett.* , 889-892, 1986.

Kim, J. M. and M. Kosterlitz, Growth in a restricted solid-on-solid model, *Phys. Rev. Letters*, 62, 2289-2292, 1989

Krug, J., Scaling relation for a growing interface, *Phys. Rev. A* , 5465-5466, 1987.

Kumar, P. and E. Foufoula-Georgiou, A multicomponent decomposition of spatial rainfall fields: 2. self-similarity in fluctuations, *Water Resour. Res.*, 29(8), 2533-2544, 1993.

Kundu, P.K. and T.L Bell, Space-time scaling behavior of rain statistics in a stochastic fractional diffusion model, *J. Hydrology*, 322, 49, 2006.

Lam, P. and F. Family, Dynamics of height-conserved surface growth model with spatially correlated noise, *Phys. Rev. A* , 7939-7950, 1991.

Liu, P. and M. Plischke, Universality in two- and three-dimensional growth and deposition models, *Phys. Rev. B* , 4781-4787, 1988.

Lovejoy, S. and D. Schertzer, Multifractals, universality classes and satellite and radar measurements of clouds and rain fields, *J. Geophys. Res.*, , 2021, 1990

Marsan, D., D. Schertzer, and S. Lovejoy, Causal space-time multifractal processes: Predictability and forecasting of rainfall fields, *J. Geophys. Res.*, **(D21)**, 26333, 1996

Meakin, P., P. Ramanlal, L. Sander and R. Ball, Ballistic deposition on surfaces, *Phys. Rev. A* , 5091-5103, 1986.

Medina, E., T. Hwa, T. Kardar and Y. Zhang, Burgers equations with correlated noise: Renormalization-group analysis and applications to directed polymers and interface growth, *Phys. Rev. A* , 3053-3075, 1989.

Over, T. M. and V. K. Gupta, A space-time theory of mesoscale rainfall using random cascades, *J. Geophys. Res.*, **(D21)**, 26319, 1996

Passalacqua, P., Porte-Agel, F., E. Foufoula-Georgiou, and C. Paola, Application of dynamic subgrid-scale concepts from large-eddy simulation to modeling landscape evolution, *Water Resour. Res.*, 42, W06D11, doi:10.1029/2006WR004879, 2006

Perica, S. and E. Foufoula-Georgiou, Linkage of scaling and thermodynamic parameters of rainfall: Results from midlatitude mesoscale convective systems, *J. Geophys. Res.*, **(D3)**, 7431-7448, 1995

Rubio, M.A., C.A. Edwards, A. Dougherty, and J.P Gollub, Self-affine fractal interfaces from immiscible displacement in porous media, *Phys. Rev. Letters*, 63, 1685-1688, 1989

Sapozhnikov, V.B. and M.G. Goldiner, Ordered phase AB diffusive growth, *Journal of Physics A: Mathematical and General*, L853, 1991.

Sapozhnikov, V.B. and M. Goldiner, Evolution of interface upon interdiffusion, *Physics Letters A* , 59-62, 1992.

Schwartz, M. and S. F. Edwards, Nonlinear deposition: a new approach, *Europhysics Letters*, 20, 301-305, 1992

Sornette, D. and Y-C. Zhang, Nonlinear Langevin model of geomorphic erosion processes, *Geophys. J. Int*, 113, 382-386, 1993

Stomfai, E., and L. M Sander, Scaling and river networks – A Landau theory for erosion, *Phys. Rev. E*, 56, R5-R8, 1997

Venugopal V. and E., Foufoula-Georgiou, Energy decomposition of rainfall in the frequency-scale domain using wavelet packets, *J. Hydrology*, 187, 3-27, 1996

Venugopal, V. E. Foufoula-Georgiou and V. Sapozhnikov, Evidence of dynamic scaling in space-time rainfall, *J. Geophys. Res.*, 104 (D24), 31599-31610, 1999

Vicsek, T., M. Cserzo, and V.K. Horváth, Self-affine growth of bacterial colonies, *Physica A*, 167, 315-321, 1990

Zhang, Y.-C, Non-universal roughening of kinetic self-affine interfaces, *Journal de Physique*, 51, 2129-2134, 1990

7 Storm Tracking and Ensemble Prediction

Lizzie S. R. Froude[1], Lennart Bengtsson[2] and Kevin I. Hodges[3]

University of Reading, Environmental Systems Science Centre, [1] lsrf@mail.nerc-essc.ac.uk, [2] olb@mail.nerc-essc.ac.uk, [3] kih@mail.nerc-essc.ac.uk

Abstract. The prediction of extratropical cyclones by the European Centre for Medium Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) Ensemble Prediction Systems (EPS) is investigated using a storm-tracking forecast verification methodology. The cyclones are identified and tracked along the forecast trajectories so that statistics can be generated to determine the rate at which the position and intensity of the forecast cyclones diverge from the corresponding analysed cyclones with forecast time. Overall the ECMWF EPS has a slightly higher level of performance than the NCEP EPS. However, in the southern hemisphere the NCEP EPS has a slightly higher level of skill for the intensity of the storms. The results from both EPS indicate a higher level of predictive skill for the position of extratropical cyclones than their intensity and show that there is a larger spread in intensity than position. The results also illustrate several benefits an EPS can offer over a deterministic forecast.

1 Introduction

Extratropical cyclones are synoptic scale (~1000 km) low pressure weather systems, which occur in the midlatitudes. They develop as instabilities in the jet stream flow in regions with large horizontal temperature gradient and vertical wind shear. A good example of such a region is the east coast of North America where there is a sharp contrast between cold polar continental air lying over the land and warm tropical air over the ocean, warmed by the poleward flow of warm water from the Gulf Stream. These transient weather systems, often referred to simply as “storms”, are characterized by circulating winds, which transport cold polar air equatorward and warm subtropical air towards the poles.

The day-to-day weather in the midlatitudes is largely dependent on the presence or absence of extratropical cyclones. In the presence of these cyclones, weather conditions are generally unsettled, stormy, wet and windy; in their absence, the weather is more settled and dry. Extratropical cyclones can be beneficial, in that they provide the majority of the precipitation received in the midlatitudes and are therefore important for

human activities such as agriculture. They can also be very damaging, since under certain conditions they can intensify more than usual, bringing very heavy rainfall and extremely strong winds. This can result in loss of life and economic damage and it is therefore important that these cyclones are predicted as accurately and far in advance as possible by numerical weather prediction (NWP).

Operational NWP models are based on a set of equations known as the "primitive equations". These models are integrated from a best estimate of the current atmospheric state, known as the analysis or the initial state, to obtain a numerical forecast of the future atmospheric state. An analysis is obtained by statistically combining different types of observations (e.g. satellite, radiosonde, aircraft) with a previously obtained numerical forecast using a process known as "data assimilation" (see for example Kalnay 2003, chapter 5). In the past NWP was deterministic in nature; a single forecast model was integrated from a single initial state. However, the potential level of forecast skill that can be provided by such an approach is limited, due to the atmosphere's chaotic behavior (Lorenz 1963). Even with a perfect forecast model, small errors in the initial conditions will grow rapidly, resulting in a total loss of predictability at higher forecast times. Today's models are clearly not perfect, and deficiencies in the models provide a further source of forecast error.

In 1992, both the European Centre for Medium Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) started to produce ensemble forecasts (Buizza and Palmer 1995; Molteni, Buizza, Palmer and Petroliagis 1996; Toth and Kalnay 1993, 1997) in addition to their deterministic products. Other operational centers have subsequently also introduced ensemble prediction systems (EPS). Ensemble prediction provides a method of extending the limit of predictability inherent in deterministic systems. By integrating an ensemble of forecasts, each started from slightly different initial conditions, an estimation of the probability density function of forecast states can be obtained (Leith 1974). One of these forecasts is integrated from the analysis and is referred to as the control forecast. The initial conditions for the other ensemble members are obtained by applying perturbations to the analysis, with the aim of sampling the probability density function of the errors in the initial state. In some EPS, perturbations are also applied to the forecast model to represent the model uncertainty (e.g. ECMWF, Buizza, Miller and Palmer 1999).

A variety of methods have been developed for generating the initial condition perturbations. Indeed the question of how to best perturb the initial conditions is a subject right at the heart of current debate (Hamill, Snyder and Morss 2000). EPS of different operational centers also differ in many other respects. They have different models, resolutions, data assimilation methods, numbers of ensemble members and model perturbations methods. It is therefore important that the forecasts generated from the different EPS are verified and compared, to explore the impact these factors have on forecast skill and to determine how the forecast systems could be improved. There are a number of standard methods used to verify EPS, such as the root mean square error, the Brier score (Brier 1950), pattern anomaly correlation (Wei and Toth 2003) and relative operating characteristics curves (Mason and Graham 2002). These forecast verification measures can be applied to any meteorological field; however, in practice they are often applied to the 500-hPa geopotential height field. For example, Buizza, Houtekamer, Toth, Pellerin, Wei and Zhu (2005) compared the performance of the

ECMWF, NCEP and Meteorological Service of Canada (MSC) EPS, using all the verification methods discussed above with the 500-hPa geopotential height field. Since these methods generally focus on entire meteorological fields, they do not provide direct information about the prediction of extratropical cyclones.

A new "storm-tracking approach" to forecast verification was presented recently in the study of Froude, Bengtsson and Hodges (2007a, hereafter FBHa) to explore the prediction of extratropical cyclones and the impact that observations of different types have on their prediction. The approach involves the identification and tracking (Hodges 1995, 1999) of extratropical cyclones along forecast trajectories. Statistics can then be generated to determine the rate at which the position, intensity and other properties of the forecast cyclones diverge from the analysed cyclones with increasing forecast time. Since extratropical cyclones are so fundamental to the weather in the mid-latitudes, the method provides a useful measure of a forecast model's ability to predict the weather. In a following study, Froude, Bengtsson and Hodges (2007b, hereafter FBHb) used the storm-tracking method to evaluate the prediction of extratropical cyclones by the ECMWF and NCEP EPS. Although, prior to this, there had been some studies of the prediction of individual extratropical cyclones by EPS (Buizza and Hollingsworth 2002; Buizza and Chessa 2002; Jung, Klinker and Uppala 2005), FBHb presented the first statistical analysis.

In this chapter, we describe the storm-tracking forecast verification method used in FBHb and present and discuss a selection of the results from this study. The chapter is organized as follows. In section 2 a description of the ECMWF and NCEP EPS data used in FBHb is given. This is followed by a description of the storm-tracking verification method in section 3. An example of the prediction of a Pacific cyclone by the two EPS is discussed in section 4 to help illustrate the storm-tracking methodology. In section 5 some of the statistical results of FBHb are presented and some conclusions are discussed in section 6.

2 Data Description

The ECMWF EPS data used in FBHb is from the time period of 6 January – 5 July 2005. During this time period the ensemble forecasts were integrated at a spectral resolution of T255 with 40 vertical levels using the operational forecast model. The ECMWF EPS consists of 50 perturbed members integrated out to 10 days at 0000 UTC and 1200 UTC. There is also a control forecast, integrated from the unperturbed analysis, but at the same resolution as the perturbed members. This analysis is obtained from the operational early-delivery 4DVAR data assimilation system (Haseler 2004). A singular vector based method is used to perturb the initial conditions in the extratropics (Buizza et al. 1995; Molteni et al. 1996). The singular vectors are computed at a horizontal resolution of T42 with 40 vertical levels using a model with simplified physics. A total energy norm (Buizza et al. 1995) is used as the measure of growth in the singular vector computation, with an optimization time of 48 hours. Both initial and evolved singular vectors are used, corresponding to those perturbations that grow fastest in the next 48 hours of the forecast and the 48 hours prior to the forecast start time respectively (Barkmeijer et al. 1999). In

addition to the initial condition perturbations, the model is perturbed by applying random perturbations to the parameterized physical processes (stochastic physics, Buizza et al. 1999).

The NCEP EPS data used in FBHb is from the time period of 6 January – 5 April 2005. During this time period the ensemble forecasts were integrated out to 16 days at a resolution of T126L28 for the first 7.5 days. From which point the resolution was reduced to T62L28 for the remaining 8.5 days of the forecast. NCEP has a smaller ensemble than ECMWF with just 10 perturbed members, but it is run more frequently, every 6 hours at 0000, 0600, 1200, 1800 UTC. There is also a control forecast integrated from the unperturbed analysis at the same resolution as the perturbed members, but only once a day at 0000 UTC. NCEP use a 3DVAR data assimilation system to generate their analyses, known as the spectral statistical interpolation (SSI) analysis system (Parrish and Derber 1992). A methodology known as the bred-vector method (Toth et al. 1993, 1997) is used to perturb the initial conditions. The method selects growing errors generated in the data assimilation process by running breeding cycles. A breeding cycle is initiated by adding/subtracting a random perturbation to/from the analysis at time t_0. The full non-linear model is then integrated forward in time from both these initial states for 24 hours to time t_1. The difference between the two forecasts is computed and rescaled using a regional rescaling algorithm (Toth et al. 1997). This difference is then added and subtracted from the analysis at time t_1 to form two bred-vector perturbations and the process is then repeated forward in time. NCEP use 5 breeding cycles to generate their initial condition perturbations and unlike ECMWF they do not perturb the forecast model.

3 Storm-Tracking Verification Approach

The extratropical cyclones were identified and tracked along the 6-hourly forecast trajectories of each of the perturbed ensemble members and the control forecasts in both hemispheres using the automated tracking scheme of Hodges (1995, 1999). Before the cyclones were identified the planetary scales with total wavenumber less than or equal to five were removed (Hoskins and Hodges 2002, 2005) so that the cyclones could be identified as extrema without being masked by the larger scales. The data were also reduced to a resolution of T42, to ensure that only the synoptic scale features were identified. Vorticity features, at the 850-hPa level (ξ_{850}), exceeding a magnitude of 1.0×10^{-5} s^{-1} were identified, as positive extrema in the NH and negative extrema in the SH, and considered as cyclones. Once the cyclones had been identified the tracking was performed, which involves the minimization of a cost function (Hodges 1999) to obtain smooth trajectories (storm tracks). Only those storm tracks that lasted at least two days, traveled further than 1000 km and had a majority of their lifecycle in 20°N - 90°N or 20°S - 90°S were retained for the statistical analysis. The identification and tracking was also performed with the ECMWF and NCEP operational analyses for the selected time periods to use for the verification.

In order to compare the forecasted cyclones with the analysed cyclones it was necessary to have an automated objective method of determining which forecast cyclones

correspond to which analysis cyclones. This was achieved using a matching methodology, modified from that of Hodges, Hoskins, Boyle and Thorncroft (2003, 2004) and Bengtsson, Hodges and Froude (2005). For the results of this chapter, a forecast track was said to match an analysis track (i.e. considered to be the same storm) if the following criteria were satisfied:

(i) At least 60% of their points overlapped in time, i.e. $100 \times (2n_m / n_A + n_F) \geq 60\%$ where n_A and n_F denote the total number of points in the analysis and forecast tracks respectively and n_m denotes the number of point in time that occur in both the analysis and forecast tracks.

(ii) The *geodesic* separation distance between the first 4 points in the forecast track, which coincide in time with the analysis track, and the corresponding points in the analysis track must be less than 4°.

The ECMWF EPS tracks were matched with the ECMWF analysis tracks and the NCEP EPS tracks were matched with the NCEP analysis tracks. The matched forecast tracks were then used to generate statistics. In the studies of FBHa and FBHb a variety of different matching criteria were considered. Although the choice of matching criteria had a considerable impact on the number of forecast tracks that match analysis tracks, the statistics generated from the matched tracks were unaffected. In this chapter we therefore only present the results obtained with the above criteria.

As an additional constraint, only those storms whose genesis occurs within the first 3 days of the forecast or that already existed at time 0 were considered. Results from the study of Bengtsson et al. (2005) indicated that the skill in predicting storm tracks after 3 days is relatively low. If a storm was generated in a forecast at a lead time greater than 3 days, and matches a storm in the analysis, then it was probably more due to chance than an accurate prediction. Although this may not be the case for the more recent forecast and analysis systems used in FBHb, we kept this constraint so that the methodology is consistent with FBHa and the statistics of two studies could be compared. This constraint will be referred to as the "genesis constraint".

The number of ensemble members that match will vary for different storms and forecast start times. In addition the storm tracks of the different ensemble members will be different lengths and so the number of ensemble member tracks available decreases with increasing forecast time. The statistics of section 5 therefore only include those data points where at least 5 perturbed member tracks are available, since calculating these values from less than 5 members would not be very informative. Restricting the diagnostics further by increasing this value of 5 does have some effect on the diagnostics, but it also limits the amount of data available, particularly at the higher forecast times. We believe that as long as all the diagnostics are produced with the same restriction, the conclusions will be unaffected by the choice of value.

4 Example Storm

Before we present some of the statistical results from FBHb, an example of an intense Pacific storm predicted by both the ECMWF and NCEP EPS is discussed to help put the statistics into context. The storm was first identified in the ECMWF analysis at 0600 UTC 1 March 2005 in southeast China (see Fig. 7.1a). It then traveled over Taiwan and across the Pacific south of Japan reaching its peak 5 days later.

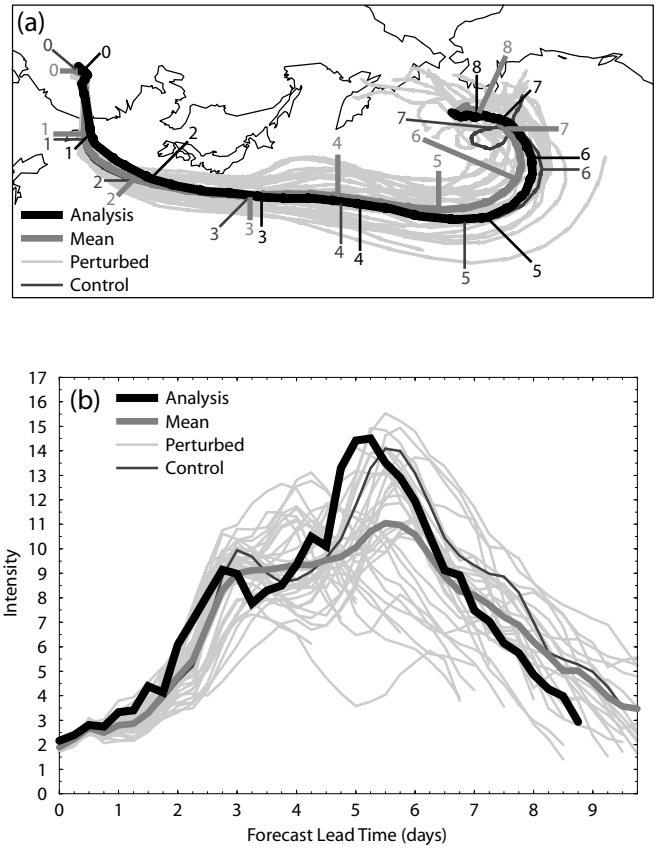


Fig. 7.1. Example of the prediction of a Pacific storm by the ECMWF EPS, reproduced from FBHb with permission of AMS. (a) Shows the analysed track and tracks predicted by the ensemble forecast started 1200 UTC 1 Mar 2005. (b) Shows the corresponding intensities in units of 10^{-5} s^{-1} (relative to the background field removal). The mean track and mean intensity of the ensemble members is also shown. The numbers labeling the tracks correspond to the forecast time in days. (A color version of this figure appears between pages 196 and 197).

The track of the storm then curved in to the northwest whilst decaying over the next 3½ days. Figure 7.1a and b shows the tracks and intensities predicted by the ECMWF ensemble forecast started from 1200 UTC 1 March. Here the control forecast predicts the track of the storm very well. The control also predicts the amplitude of the storm very well; it is just slightly out of phase with the analysis. Thirty-seven of the perturbed members predict the storm with the chosen matching criteria (see section 3). The ensemble member tracks are evenly distributed about the analysed track, resulting in a high quality mean track prediction that is better than the control forecast at the high forecast times. It can also be seen from the numbers indicating the forecast time on the mean track and analysed track that the ensemble member storms

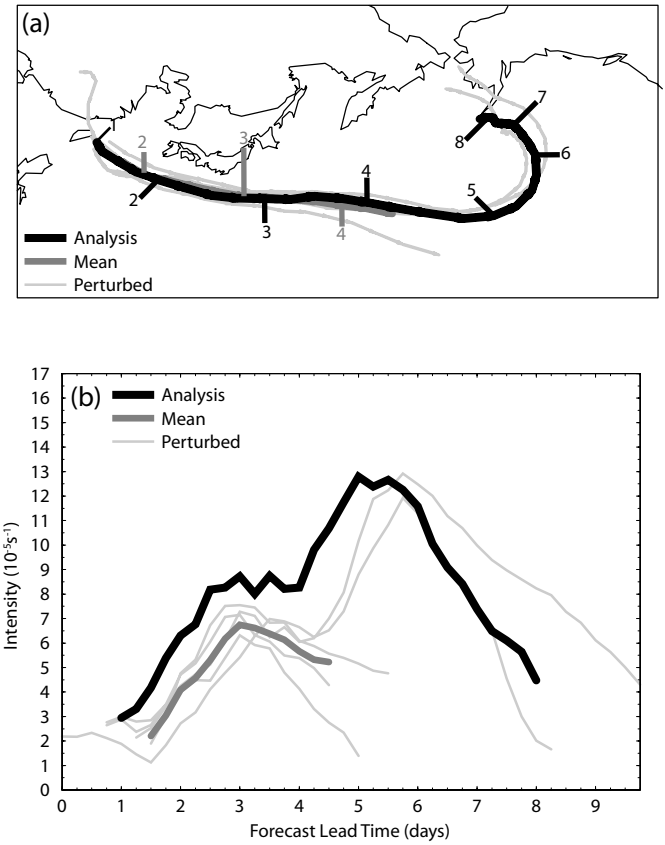


Fig. 7.2. As in Fig. 7.1 but for the NCEP EPS, reproduced from FBHb with permission of AMS. The mean track and intensity is shown when there are 5 ensemble members available. A NCEP control forecast was not available for this forecast start time. (A color version of this figure appears between pages 196 and 197).

are generally moving at a slower speed than the analysed storm. Some of the perturbed members, like the control, predict the double-peaked shape of the analysis, whilst others reach their peak a day or more early and then completely decay without first regaining in intensity. The mean intensity significantly underpredicts the storms amplitude, between day 4 and 6, around the storms peak.

Figure 7.2a and b shows the tracks and intensity of the storm predicted by six ensemble members of the NCEP EPS started at the same time as the ECMWF forecast of Fig. 7.1 (1200 UTC 1 March). The storm is not identified in the analysis until 1200 UTC 2 March, 30 hours after the ECMWF analysis and it does not reach as high a peak in intensity. This is probably due to the lower resolution of the NCEP analysis system. The mean track and intensity is shown when there are at least 5 ensemble members available to average. Two of the ensemble members predict the track and intensity of the storm very well, but the other four members do not regain intensity after the initial growth and have decayed completely by day 5 or 6 of the forecast. The ensemble members underpredict the amplitude of the cyclone. As with the ECMWF EPS, the forecasted cyclones are propagating at a slower speed than the analysed cyclone on average.

5 Statistical Results

5.1 Percentage of forecast storm tracks that match

In this section we present a selection of the statistical results of FBHb. Since the statistics are generated from *only* those forecast storm tracks that match, it is important to consider the number of tracks that match before interpreting the statistics. Table 7.1 shows the percentage of forecast tracks from all the perturbed ensemble members that match with analysis tracks, for the ECMWF and NCEP EPS. It also shows the percentage of tracks from just the control forecast that match with analysis tracks. The numbers in brackets are the total number of forecast tracks, which satisfy the genesis constraint (see section 3) and the percentage values are the percentage of these forecast tracks that match with analysis tracks.

A higher percentage, of the perturbed ensemble member tracks and the control forecast tracks, match for the ECMWF EPS than for the NCEP EPS. The difference between the percentage of ECMWF and NCEP perturbed member tracks that match is comparable to the difference between the percentage of ECMWF and NCEP control forecast tracks that match. This suggests that the superior skill of the ECMWF EPS is due to the higher resolution model and 4DVAR data assimilation system rather than the perturbation methodology. Another noticeable observation is that the percentage of control forecast tracks that match is higher than that of the perturbed member tracks for both the ECMWF and NCEP EPS. Since the control forecast has been generated from an optimal analysis, one might expect it to be better than the perturbed members for the earlier part of the forecast. The spatial matching focuses

Table 7.1. Percentage of ECMWF and NCEP perturbed and control ensemble member tracks that match with analysis tracks in the NH and SH, adapted from FBHb with permission of AMS. The numbers in brackets are the total number of forecast tracks that satisfy the genesis constraint. The percentage values correspond to the percentage of these tracks that match analysis tracks.

	Percentage Match	
	ECMWF	NCEP
NH:		
Perturbed	38.9 (414762)	36.5 (83928)
Control	47.5 (8239)	42.6 (2089)
SH:		
Perturbed	39.7 (432389)	34.3 (85201)
Control	48.8 (8683)	42.3 (2120)

on the earlier parts of the forecast tracks and this may therefore be the reason why a higher percentage of control forecast tracks match.

In FBHb the percentage of forecast tracks that match was considered for three different matching criteria. Whilst the percentage match varied considerably for the different criteria, the difference in percentages between the ECMWF and NCEP EPS, and the perturbed and control ensemble members was consistent. As mentioned previously the statistics generated from the matched tracks were unaffected by the choice of criteria. We now present some statistical results generated from the matched tracks.

5.2 Forecast skill of perturbed members and control

Figure 7.3 shows some diagnostics for the cyclones predicted by the ECMWF EPS. The solid lines of Fig. 7.3a and c show, in the NH and SH respectively, the minimum, mean and maximum separation distance of the matched perturbed member storm tracks from the analysis tracks as a function of forecast time. It should be noted that the mean curve is the mean separation distance of the perturbed members from the analysis and not the separation distance of the mean track from the analysis, which is investigated in the next subsection. The mean curve in the Figure. therefore corresponds to the average skill of the perturbed member forecasts.

The dotted lines of Fig. 7.3a and c show the separation distance of the best and worst track of the ensemble from the analysis. To calculate the best and worst track the average separation distance of each ensemble member track from the corresponding analysis track over their whole lifetime was computed. The ensemble members with the lowest and highest values were taken to be the best and worst tracks respectively. The minimum (and maximum) error of the ensemble will be obtained from different ensemble member tracks at different forecast times, whereas the best (and worst) ensemble member error is obtained from the same ensemble member at all

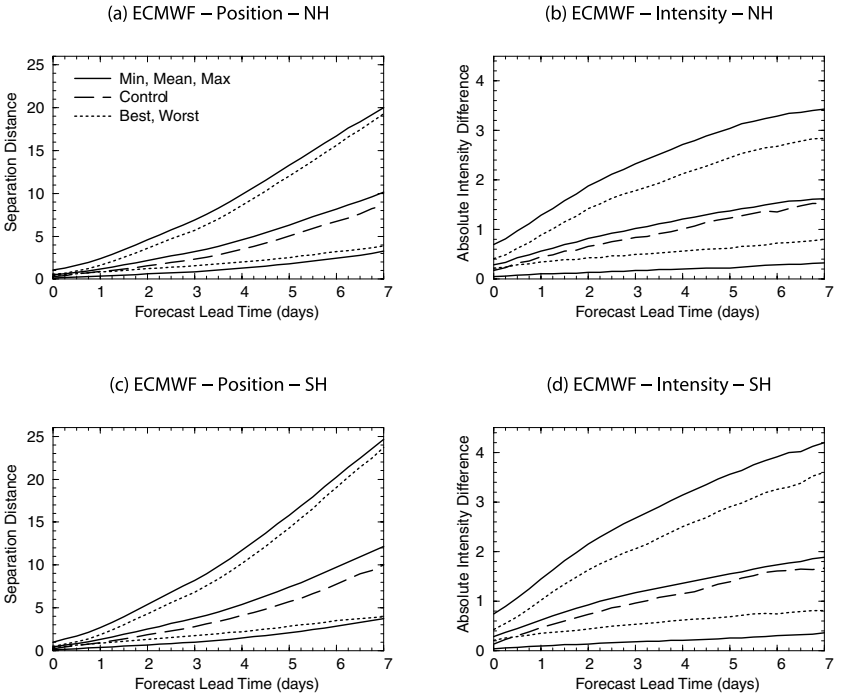


Fig. 7.3. Ensemble diagnostics for the ECMWF EPS, reproduced from FBHb with permission of AMS. The solid lines show the minimum, mean and maximum separation distance/absolute intensity difference between the matched perturbed member tracks and analysis tracks as a function of forecast time. The dotted lines show the separation distance/absolute intensity difference between the best and worst perturbed ensemble members (for details see text) and analysis tracks. The dashed lines show the separation distance/absolute intensity difference between the control forecast tracks and the analysis tracks. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to the background field removal) respectively.

forecast times. The diagnostics described above have included the perturbed ensemble members only and not the control. In Fig. 7.3a and c the dashed line shows the separation distance of the matched control forecast tracks from the analysis tracks. Panels b and d of Fig. 7.3 show the same diagnostics, but for absolute intensity difference rather than separation distance. Here the best and worse ensemble member tracks are determined by the average absolute intensity difference, between forecast and analysis tracks, over their whole lifetime. Hence the best ensemble member for the separation distance results is not necessarily the best member for the intensity results. In fact calculations show that only about 12% of ensemble members that are best for position are also best for intensity. Figure 7.4 shows the same diagnostics for

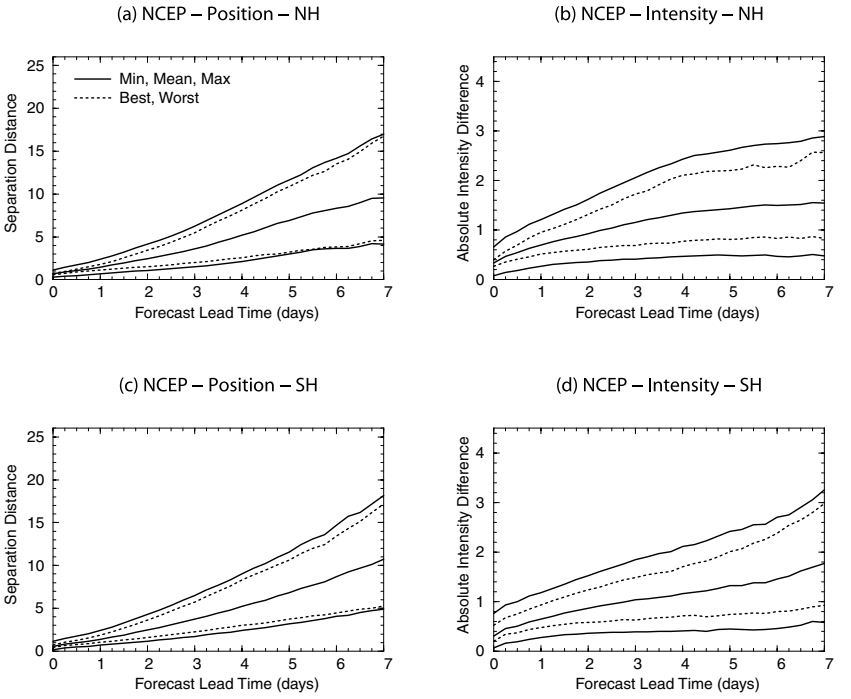


Fig. 7.4. As for Fig. 7.3 but for the NCEP EPS, reproduced from FBHb with permission of AMS. The diagnostics for the control forecast are not included.

the NCEP EPS. Since the NCEP control forecast is only run once a day, the NCEP EPS data we currently have available is not sufficient for producing the diagnostics for the control forecast.

By comparing the separation distance curves with the intensity curves, it can be seen that the error growth rates of the storms intensity differs considerably to that of the storms position for both the ECMWF and NCEP EPS. For intensity the errors grow rapidly in the earlier part of the forecast, but at higher lead times the errors grow at a slower rate and the curves become almost flat showing signs of saturation. The error in position, on the other hand, grows quite slowly at the beginning of the forecast, but becomes faster at the higher lead times. This difference between the error growth rates suggests a higher level of skill in the prediction of the position of the cyclones than their intensity. Initially it was thought that the spatial matching may have some impact on the growth of the separation distances. However, in FBHa different spatial matching methods were investigated extensively and found to have no significant effect on the results. There is also a larger spread (more uncertainty) in intensity than position. This is deduced by comparing the minimum and maximum curves with the mean curves. For example for the ECMWF EPS in the NH (Fig. 7.3a and b) the maximum error in position at day 3 is about 7° and the mean error reaches this value about 2½ days

further into the forecast. Similarly the mean error at day 3 is about 3° and the minimum error curve reaches this value about 3½ days further into the forecast. Now considering the intensity curves, the maximum error at day 3 is about 2.4×10^{-5} s^{-1}, but the mean error does not reach this value in the 7 day forecast range plotted. A similar result is found when comparing the minimum and mean intensity errors.

The ECMWF control forecast has consistently higher skill than the perturbed members by ½ to 1 day for both position and intensity. This is to be expected, in the earlier part of the forecast, since the initial part of the control forecast has been optimized by the 4DVAR data assimilation system. If the perturbed ensemble members were obtained using only initial condition perturbations (and not model perturbations), then it might be expected that the error of the control forecast would converge to that of the perturbed members at higher lead times. However, the ECMWF EPS also includes model perturbations, which may have an impact on the skill of the perturbed members at higher lead times. Unfortunately we have insufficient data for the NCEP control forecast to determine whether it is also consistently better than the perturbed members.

The difference between the mean curves and the best ensemble member curves is significant showing that the best ensemble member can provide a much better prediction of a cyclone than a single deterministic forecast. For the ECMWF EPS the day 5 skill of the best ensemble member is comparable to the day 3 skill of the control and to the day 2½ skill of the mean curve for the position of the cyclones. For the intensity of the storms these improvements are even greater, increasing by about 1 day. Similar improvements are also gained from the best ensemble member of the NCEP EPS. The high level of skill of the best ensemble member is encouraging in itself because it indicates that the errors in the initial state are being sampled effectively. However, from a practical point of view the question of how soon into the forecast the best ensemble member can be determined is more important. If, for example, the ensemble member that is best for the first day of the forecast is still best (or better than the average ensemble member) some time further into the forecast then this would provide helpful information to an operational weather forecaster. This was investigated by selecting the ensemble member that was best for the first day and first two days of the forecast and then computing the error growth of this ensemble member (not shown). Rather disappointingly the error of these selected ensemble members diverges very quickly to that of the average ensemble member.

The perturbed members and the control forecast of the ECMWF EPS have slightly less skill in the SH than the NH. A more noticeable feature is the difference between the ensemble minimum and maximum, which is considerably larger in the SH than the NH. The NCEP perturbed members also have slightly less skill in the SH than the NH, but the difference between the minimum and maximum is very similar in the two hemispheres.

Comparing the diagnostics of Figs. 7.3 and 7.4 is not a very fair or objective comparison, since the NCEP EPS consists of a much smaller number of ensemble members than the ECMWF EPS. In Fig. 7.5 the minimum, mean and maximum diagnostics are shown for the NCEP EPS and for a 10 member version of the ECMWF EPS, obtained by randomly selecting 10 of the 50 ECMWF ensemble members. In the

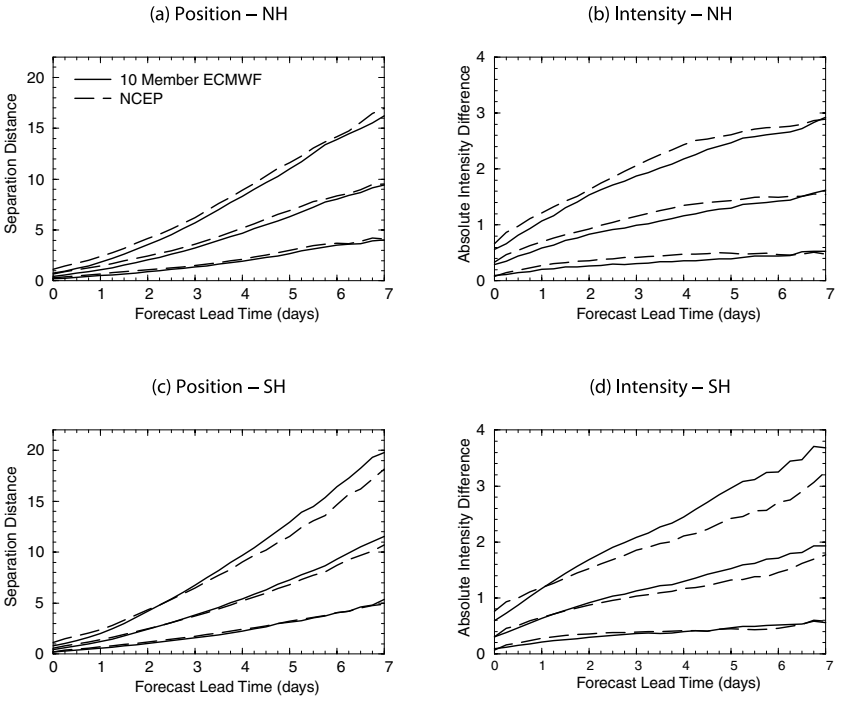


Fig. 7.5. Minimum, mean and maximum separation distance/absolute intensity difference between the matched perturbed member tracks and analysis tracks for the NCEP EPS (dashed lines) and for 10 randomly selected perturbed members of the ECMWF EPS (solid lines) as a function of forecast time. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to the background field removal) respectively. Reproduced from FBHb with permission of AMS.

NH, the ECMWF perturbed ensemble members have approximately ½ a day more skill in predicting the position of the cyclones, and approximately 1 day more skill in predicting the intensity of the cyclones, than the NCEP perturbed ensemble members. However in the SH, the NCEP perturbed ensemble members are about ½ a day better, from day 4 of the forecast, at predicting the position of the cyclones and are about 1 day better, from day 3 of the forecast, at predicting the intensity. As with the full 50 member ECMWF EPS the spread of the 10 member ensemble is larger in the SH than the NH. This is investigated in more detail in the next subsection.

5.3 The ensemble mean and spread

One of the aims of ensemble prediction is that the average of the ensemble forecasts will provide a forecast that is, although somewhat smoothed, superior to the control forecast (Leith 1974; Toth et al. 1993, 1997). This is investigated in this subsection.

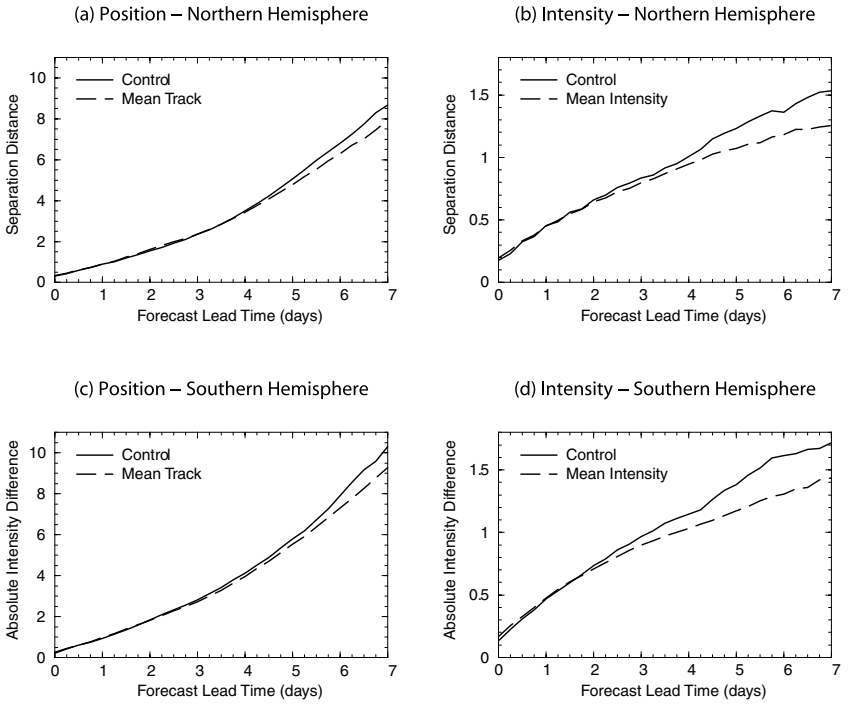


Fig. 7.6. Separation distance/absolute intensity difference between ECMWF ensemble mean tracks (computed from the matched perturbed members and the control) and analysis tracks (dashed lines), and ECMWF control forecast tracks and analysis tracks (solid lines) as a function of forecast time. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to the background field removal) respectively. Reproduced from FBHb with permission of AMS.

The results are only presented for the ECMWF EPS, since we have insufficient data at this time to generate the diagnostics for the NCEP EPS. For each ensemble forecast the mean track and mean intensity of the ensemble member tracks (including the control) that matched were computed. Figure 7.6a and c shows the mean separation distance of the mean tracks and control tracks from the analysis tracks in the NH and SH respectively. The skill of the mean track and control track is almost identical until day 4 of the forecast, from which point the error growth of the control track becomes slightly larger. By day 7 of the forecast the mean track has about ½ a day advantage over the control forecast in both hemispheres. However, this may be of little benefit, since forecasts of this high lead time will have low levels of skill in general.

Figure 7.6b and d shows the results for absolute intensity difference. The difference between the error growth rates of the ensemble mean and control is more significant for the cyclone intensity than for position. From day 2 of the forecast the error growth of the control forecast is larger than that of the ensemble mean and by day 7 of the forecast the ensemble mean has about 2 days more skill than the control forecast. The skill of the ensemble mean is higher in the NH than the SH for both position and intensity of the cyclones.

For an EPS to be statistically reliable the average distance of the ensemble mean from the analysis should be equal to the average distance of the ensemble members from the ensemble mean, i.e. the error of the ensemble mean should be equal to the ensemble spread. Ensemble spread was calculated as the mean separation distance/absolute intensity difference of the ensemble member tracks from the ensemble mean track. Figure 7.7 shows the ensemble spread and ensemble mean skill for the

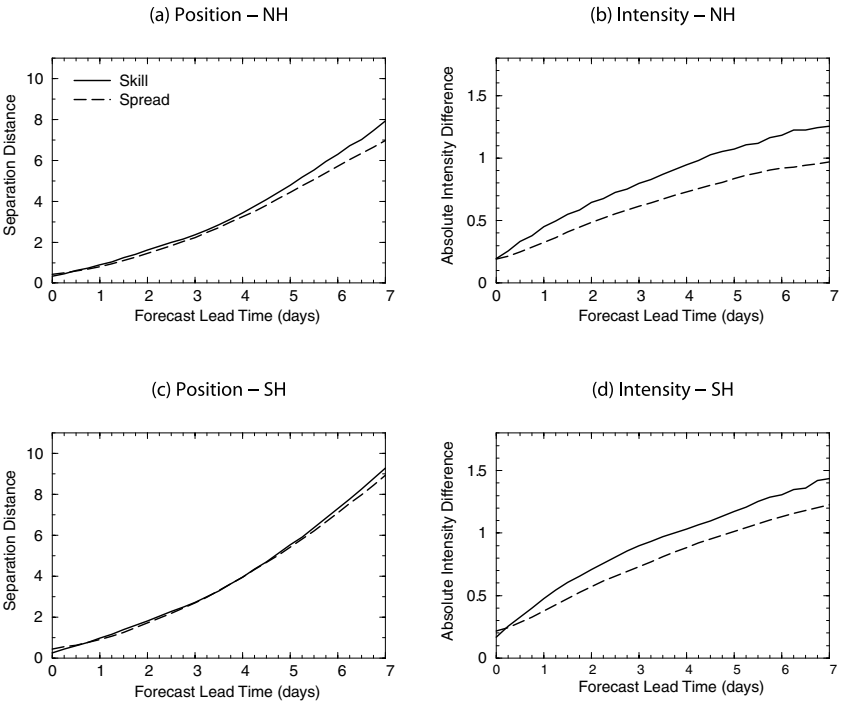


Fig. 7.7. Spread skill diagnostics for ECMWF EPS, reproduced from FBHb with permission of AMS. The solid lines show the mean skill and are the same as the dashed lines of Fig. 7.6. The dashed lines show the spread calculated as the mean separation distance/absolute intensity difference between the matched perturbed member tracks and the ensemble mean tracks. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to the background field removal) respectively.

position and intensity of the cyclones in both hemispheres. The difference between the ensemble mean error and the ensemble spread is very small for the position of the cyclones. In the NH the spread is almost identical to the skill until day 4 when the ensemble mean error begins to exceed the spread ever so slightly. In the SH the results are even more encouraging, since very little difference can be seen between the two curves. The lower level of skill in the SH corresponds well with the larger spread. There is a much larger difference between the mean error and spread curves for the intensity of the cyclones. The error growth of the ensemble mean is larger than the spread in both hemispheres, but there is more of a difference in the NH. This shows that the intensities predicted by the ensemble members are not evenly distributed about the analysed intensities. This under-dispersion of the ECMWF EPS was also found by Buizza et al. (2005), for the 500-hPa geopotential height field, from about day 5 of the forecast. The Buizza et al. (2005) study also found this to be the case, and to a greater extent, for the NCEP and MSC EPS.

6 Conclusions

In this chapter a selection of results from FBHb have been presented and discussed. The FBHb study used a storm-tracking forecast verification approach, first introduced in FBHa, to explore the prediction of extratropical cyclones by the ECMWF and NCEP EPS. Whilst the storm-tracking approach is more difficult and time consuming to implement than more conventional verification methods, it provides very useful information about the prediction of storms that cannot be obtained from the traditional approaches. As with any verification method the storm-tracking method has limitations and biases, which are discussed in FBHa.

Overall the ECMWF EPS had a slightly higher level of performance than the NCEP EPS, but the results suggest that this is more to do with the model and data assimilation system than the perturbation methodology. This agrees with the results of the Buizza et al. (2005) study, which used the more conventional methods of ensemble verification discussed in the introduction to compare the ECMWF, NCEP and MSC EPS. The large number of differences between EPS of different centers (e.g. resolution, model, data assimilation methods) make it difficult to compare the different perturbation methodologies. Indeed one of the main conclusions of the Buizza et al. (2005) study was the difficulty in comparing such different systems and that further studies, in which different perturbation methods are explored using a single analysis and forecast system, are necessary to determine the advantages and disadvantages of the different perturbation methodologies.

Although it has not been possible to draw any conclusions concerning the different perturbation methodologies, the results illustrate some of the benefits an ensemble forecast can offer over a single deterministic forecast. Firstly we find the high level of skill of the best ensemble member very encouraging. Clearly an ensemble forecast will always have a best ensemble member, but it is the difference in skill between this best ensemble member and that of the average member we find encouraging. One of the properties of a perfect ensemble forecast must be that one of the ensemble members

provides a perfect forecast (Buizza 1997). The skill of the best ensemble member therefore gives an indication of how close to satisfying this criterion an EPS is. It also suggests that the errors in the initial state are being sampled effectively. These arguments do however take a theoretical viewpoint. From a practical perspective, the main question is the potential to identify the best ensemble member at some useful time. Preliminary results suggest this may be very difficult.

Another benefit of EPS, the results presented in this chapter have illustrated, is the measure of the predictability of the atmosphere provided by the ensemble spread. For the position of storms, the ECMWF ensemble mean error is very close to the ensemble spread. This has a very useful practical application, since the spread of the ensemble can give the forecaster an indication of the accuracy of the ensemble mean forecast. Unfortunately the same is not true for the intensity of storms; the ensemble spread is less than the ensemble mean error.

An interesting result is the superior quality of the ECMWF control forecast to the perturbed members. It is ½ to 1 day better throughout the forecast. Unfortunately data limitations have stopped us from determining if this is also the case for the NCEP EPS. The ECMWF control forecast has been produced from the best estimate of the initial state and forecast model and may therefore be expected to have a higher level of skill than the perturbed members. We believe that the difference in skill between the control and the perturbed members should be considered as part of the evaluation of an EPS. If the control forecast has a consistently higher level of performance than the perturbed members it should perhaps be weighted accordingly in the calculation of the ensemble mean.

The results presented in this chapter indicate a higher level of predictive skill for the position of extratropical cyclones than their intensity. More detailed analysis of the position error in FBHb showed that the errors in storm propagation speed play a larger role than errors in direction. There is a small bias for forecasted cyclones to propagate too slowly and preliminary results indicate that this may correspond to a small bias for the forecast intensity to be too high. We believe that the main reason for the lower levels of skill in cyclone intensity and propagation speed than that of the direction is due to errors in the storms vertical structure. A vertical tilt is critical to the growth of extratropical cyclones, as it allows upper and lower level disturbances to interact in the feed-back process of self-development (see for example Holton 2004, chapter 8). Errors in the storms vertical structure will cause errors in both the intensity and propagation speed of the storm. In contrast the direction of the storm will be mainly determined by the 700-hPa level winds and will not be significantly affected by errors in the vertical structure. It is possible that a higher resolution and/or frequency of upper air observations than currently available would be required to accurately predict the growth and development of extratropical cyclones. The use of targeted observations (e.g. Leutbecher, Barkmeijer, Palmer and Thorpe 2002) may improve the prediction of cyclone intensity considerably.

As discussed by FBHb, the results presented in this chapter could potentially be useful to both the developers of ensemble forecast systems and to the users of these systems. A developer could use the statistical approach to evaluate the impact of future upgrades to an ensemble forecast system. The user could use the storm tracking method to help with the forecasting of individual storms. Modifications to the methodology

would be required, since the ensemble mean and spread results of this chapter are computed from ensemble member tracks that match with analysis tracks, they are unknown before the forecast verification time has past. It is suggested that in an operational forecast situation, measures of ensemble mean and spread could instead be obtained by matching perturbed ensemble member tracks with control forecast tracks, rather than with analysis tracks. Since the statistics of FBHb generated from the matched tracks were unaffected by the choice of matching criteria, we believe this could provide a useful practical measure of ensemble mean and spread.

7 Acknowledgements

We thank ECMWF and NCEP for making their ensemble forecast and analysis data available to us.

References

Bengtsson, L., Hodges, K. I. and Froude, L. S. R. (2005) Global observations and forecast skill. Tellus 57A, 512-527.

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. Mon. Wea. Rev. 78, 1-3.

Buizza, R. (1997) Potential forecast skill of ensemble prediction and spread skill distributions of the ECMWF ensemble prediction system. Mon. Wea. Rev. 125, 99-119.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. (2005) A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. Mon. Wea. Rev. 133, 1076-1097.

Buizza, R. and Chessa, P. (2002) Prediction of the U. S. storm of 24-26 January 2000 with the ECMWF ensemble prediction system. Mon. Wea. Rev. 130, 1531-1551.

Buizza, R. and Hollingsworth, A. (2002) Storm prediction over Europe using the ECMWF ensemble prediction system. Meteorol. Appl. 9, 289-305.

Buizza, R.., Miller, M. and Palmer, T. N. (1999) Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Quart. J. Roy. Meteor. Soc. 125, 2887-2908.

Buizza, R. and Palmer, T. N. (1995) The singular-vector structure of the atmospheric global circulation. J. Atmos. Sci. 52, 1434-1456.

Froude, L. S. R., Bengtsson, L. and Hodges, K. I. (2007a) The predictability of extratropical storm tracks and the sensitivity of their prediction to the observing system. Mon. Wea. Rev. 135, 315-333.

Froude, L. S. R., Bengtsson, L. and Hodges, K. I. (2007b) The prediction of extratropical storm tracks by the ECMWF and NCEP ensemble prediction systems. Mon. Wea. Rev. accepted.

Hamill, T. M., Snyder, C. and Morss, R. E. (2000) A comparison of probabilistic forecasts from bred, singular-vector and perturbed observation ensembles. Mon. Wea. Rev. 128, 1835-1851.

Hodges, K. I. (1995) Feature tracking on the unit sphere. Mon. Wea. Rev. 123, 3458-3465.

Hodges, K. I. (1999) Adaptive constraints for feature tracking. Mon. Wea. Rev. 127, 1362-1373.

Hodges, K. I.. Hoskins, B. J., Boyle, J. and Thorncroft, C. (2003) A comparison of recent reanalysis data sets using objective feature tracking: storm tracks and easterly waves. Mon. Wea. Rev. 131, 2012-2037.

Hodges, K. I.. Hoskins, B. J., Boyle, J. and Thorncroft, C. (2004) Corrigendum to "A comparison of recent reanalysis data sets using objective feature tracking: storm tracks and easterly waves". Mon. Wea. Rev. 132, 1325-1327.

Holton, J. R. (2004) *An Introduction to Dynamic Meteorology*. Elsevier Academic Press, 4th edition.

Hoskins, B. J. and Hodges, K. I. (2002) New perspectives on northern hemisphere winter storm tracks. J. Atmos. Sci. 59, 1041-1061.

Hoskins, B. J. and Hodges, K. I. (2005) A new perspective on southern hemisphere storm tracks. J. Climate. 18, 4108-4129.

Jung, T., Klinker, E. and Uppala, S. (2005) Reanalysis and reforecast of three major European storms of the twentieth century using the ECMWF forecasting system: Part II: Ensemble forecasts. Meteorol. App. 12, 111-122.

Kalnay. E. (2003) Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press.

Leith, C. E. (1974) Theoretical skill of Monte-Carlo forecasts. Mon. Wea. Rev. 102, 409-418.

Leutbecher, M., Barkmeijer, J., Palmer, T. N. and Thorpe, A. J. (2002) Potential improvements to forecasts of two severe storms using targeted observations. Quart. J. Roy. Meteor. Soc. 128, 1641-1670.

Lorenz, E. N. (1963) Deterministic nonperiodic flow. J. Atmos. Sci. 20, 130-141.

Mason, S. J. and Graham, N. E. (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quart. J. Roy. Meteor. Soc. 128, 2145-2166.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996) The ECMWF ensemble prediction system: Methodology and validation. Quart. J. Roy. Meteor. Soc. 122, 73-119.

Toth, Z. and Kalnay E. (1993) Ensemble forecasting at the NMC: The generation of perturbations. Bull. Amer. Meteor. Soc. 74, 2317-2330.

Toth, Z. and Kalnay E. (1997) Ensemble forecasting at NCEP and the breeding method. Mon. Wea. Rev. 125, 3297-3319.

Wei, M. and Toth, Z. (2003) A new measure of ensemble performance: Perturbation versus error correlation analysis (PECA). Mon. Wea. Rev. 131, 1549-1565.

8 Towards a Nonlinear Geophysical Theory of Floods in River Networks: An Overview of 20 Years of Progress

Vijay K. Gupta,[1] Brent M. Troutman,[2] and David R. Dawdy[3]

[1] University of Colorado, Dept. of Civil, Environmental and Architectural Engineering, Cooperative Institute for Research in Environmental Sciences, guptav@cires.colorado.edu
[2] U. S. Geological Survey, troutman@usgs.gov
[3] Hydrologic Consultant, dawdy1926@sbcglobal.net

Abstract. Key results in the last 20 years have established the theoretical and observational foundations for developing a new nonlinear geophysical theory of floods in river basins. This theory, henceforth called the scaling theory, has the explicit goal to link the physics of runoff generating processes with spatial power-law statistical relations between floods and drainage areas across multiple scales of space and time. Published results have shown that the spatial power law statistical relations emerge asymptotically from conservation equations and physical processes as drainage area goes to infinity. These results have led to a key hypothesis that the physical basis of power laws in floods has its origin in the self-similarity (self-affinity) of channel networks. Research within the last 20 years has also shown that self-similarity is the basis for the widely observed fractal structure and Horton relations in river networks. Observed power laws in floods span a broad range of spatial scales and multiple time scales that range from hours of individual flood events to annual time scale of flood frequencies. They serve as the foundation for developing a new diagnostic framework to test different assumptions governing spatial variability in physical processes that arise in predicting power law statistical relations between floods and drainage areas. The structure of the diagnostic framework is illustrated using two examples. Contemporary relevance and future applications of the scaling theory come from predicting floods in the context of a significant warming of the Earth's climate, which is altering local, regional, continental and global balances of water and energy. This anthropogenic perturbation to the planetary hydro-climate precludes making flood predictions at gauged and ungauged sites based on regional statistics from historical stream flow data, which is a well-established practice in hydrologic engineering.

1 Introduction

The notion of scale has assumed a central position in hydrologic science, as it is well recognized that processes and fluxes involve interactions occurring over a wide range of space and time scales (National Research Council 1991, p. 59). Dooge (1997) articulated the fundamental "challenge of scale" in hydrologic sciences. He illustrated that the spatial scales in hydrology range from 10^{-8} m for molecular clusters to 10^{7} m for the planetary scale, and time scales range from 10^{-13} seconds for

water clusters to 10^9 years for stellar evolution. The problem of scale is also intimately tied into the level of the conceptual framework at which hydraulic and hydrologic processes are modeled (Klemes 1978). The present theory spans a fairly broad subset of these space-time scales pertaining to floods in drainage basins.

Power laws, well known as scaling relations (Barenblatt 1996), have been used extensively, not only in hydrology but in many other sciences as well, to describe relationships between variables that range over multiple scales of space and time. It has not always been widely appreciated, however, that such power laws often possess a deeper significance, beyond their usefulness for fitting data. Power laws arise in scale-invariant systems that exhibit self-similarity or self-affinity, and these properties reflect a fundamental symmetry of nature (Barenblatt 1996; Schroeder 1991). Power laws arising out of self-similarity have been observed in many data sets that include geophysical, biological and social sciences (Turcotte 1997; Turcotte and Rundle 2002; West and Brown 2004).

The U. S. Geological Survey (USGS) developed the regional (spatial) quantile regression approach in the 1960s for predicting annual flood frequencies at ungauged locations in river basins where no stream flow data are available. The quantile regression approach uses historical streamflow records and the assumption of stationarity in these records for predicting flood quantiles in ungauged basins. The regression relations developed typically express quantiles as power laws involving one or more explanatory variables that pertain to drainage basin and climatic characteristics. Drainage area is usually the most important explanatory variable, and many times the only variable, used. In many empirical studies of regional flood frequencies, power law relationships have been observed to hold between annual flood statistics and drainage areas (Gupta and Dawdy 1995; Robinson and Sivapalan 1997; Eaton, Church, and Ham 2002).

The fundamental premise of this paper is that a physical understanding of the spatial power law statistics that are observed in flood regionalization studies may be developed by an appeal to scale invariance. Gupta, Castro, and Over (1996) published the first set of results to understand the physical origins of power laws or scaling relations in floods. They showed that floods exhibit power laws with respect to drainage area A under highly idealized conditions in the self-similar Peano network as $A \to \infty$. Gupta and Waymire (1998a and 1998b) generalized some of the results, and published two overview papers. These papers provided a self-contained overview of the key results that had been obtained until a decade ago, one focusing on the physical aspects and the other on the mathematical aspects. Research since then has shown that many of the results reviewed by Gupta et al. (1998a and 1998b) hold under realistic physical conditions, which are providing the theoretical and observational foundations for developing a nonlinear geophysical theory of floods in river basins. The focus of this paper is to give an overview of the recent key results, current work and future challenges.

Drainage basins in many parts of the world are ungauged or poorly gauged, and in some cases existing measurement networks are declining. The problem of prediction in ungauged basins (PUB) has been always an issue of great practical significance in engineering hydrology, especially for smaller basins, because approximately one streamflow gauging station exists for every 500 km^2 in the United States (US) (NRC

1988, p. 120). The density of gauged stream data is much worse in developing countries of the world. However, PUB is acquiring a compelling new significance due to climate change as a result of global warming. One expects greater evapotranspiration with an increase in temperature, which should produce greater precipitation and cloud cover. Indeed, Hobbins, Ramirez, and Brown (2004) used hydrological data to demonstrate that actual evapotranspiration has increased in the conterminous US in the last half a century. Global climate models (GCMs) also show an increase in evapotranspiration (Manabe 1997) and in precipitation and cloud cover (Karl, Knight, Easterling, and Quayle 1996). Indeed, major changes in the global water cycle have been observed in the last five decades in terms of gradual increase in severe weather and hydrologic events such as major hurricanes, floods and droughts, which have led to a steady increase in economic losses (IPCC 2001). One can infer a strong signature of non-stationarity in extreme weather-related and hydrologic phenomena.

The scientific consensus has grown that global warming is real and in substantial part is caused by human activities. A non-stationary climate change is resulting in the alteration of the global water cycle, and in the frequency of hydrologic extremes worldwide, because the global water cycle is coupled with climate in a complex manner involving non-linear feedbacks (Nordstrom, Gupta, and Chase 2005). Non-stationary changes in the distribution of water present an enormous global problem for the management of water resources in all countries of the world, because historical hydrologic observations, including precipitation and stream flows, no longer represent the future. In this sense, all basins become ungauged, and the PUB problem emerges as a central challenge for research in hydrologic sciences and many other related disciplines, including applications in hydrologic engineering. Klemes (1989) made compelling arguments to underscore the issue of non-stationarity in hydrologic data, and its implications for predicting probabilities of extreme floods. The International Association of Hydrologic Sciences (IAHS) has embraced PUB through a new decade-long initiative (Sivapalan, Takeuchi, Franks, Gupta, Karambiri, Lakshmi, Liang, McDonnell, Mendiondo, O'Connell, Oki, Pomeroy, Schertzer, Uhlenbrook, and Zehe 2003). The scaling theory of floods reviewed here provides the scientific foundations for making flood predictions under a changing hydro-climate, which is altering local, regional, continental and global balances of water and energy, and frequency of hydrologic extremes.

Our paper is organized as follows. Section 2 gives a very short overview of the key results that Gupta et al. (1998a and 1998b) included in the two review papers. Section 3 describes the dynamical basis of flood producing processes in terms of an equation of continuity for a channel network that Gupta et al. (1998a, 1998b) introduced. It is based on the assumption that a basin can be naturally partitioned into a large number of hillslope-link pairs that exist in a terrain for which digital elevation models (DEMs) are being used (Band 1986; Mantilla and Gupta 2005). We explain the challenge of parameterizing physical processes at the hillslope-link scale from three-dimensional representation of continuum equations governing infiltration, surface flow and unsaturated-saturated sub-surface flow in hillslopes, and stream flows in channels. Key results are summarized for Peano and Mandelbrot-Vicsek networks, which show that scaling in floods holds under more realistic physical

conditions than Gupta et al. (1998a, 1998b) considered. These results further confirmed that power law relations between floods and drainage area A emerge from physical processes as $A \to \infty$. Asymptotic appearance of power laws in highly idealized Peano and Mandelbrot-Vicsek networks clearly illustrated that a physical basis of power laws in floods has its origin in the self-similarity of river networks.

For the past five years, the main focus of our research has been to generalize the scaling theory of floods from idealized basins to real basins using real data that includes rainfall and stream flows. Observed power laws in floods for individual rainfall-runoff (RF-RO) events, and how they may be related to annual flood frequencies, are described in Section 4. They furnish the basis to develop an innovative new diagnostic framework for testing various physical assumptions for predicting power law statistics in floods from physical processes that is also needed to solve the problem of dynamic parametric complexity (Gupta 2004). The structure of the diagnostic framework is explained in Section 4. In Section 5, we include a short review of 20 years of progress in understanding self-similarity (self-affinity) in real networks and its applications to floods. A brief discussion of future challenges is given in Section 6, which includes how scaling framework provides a new research direction to make flood predictions at multiple space and time scales under a changing global hydroclimate due to global warming.

2 A Brief Overview of Early Developments in the Scaling Theory of Floods: From Engineering to Nonlinear Geophysics

Existence of sparse stream flow data sets in space and time greatly influenced the development of hydrologic models, both statistical and deterministic, during the twentieth century (National Research Council 1988, p. 120). In the 1960s, the USGS developed the quantile regression method for predicting annual flood frequencies at ungauged basins in "homogeneous spatial regions". The idea of homogeneous regions enabled the USGS to pool streamflow data from different basins in the same region, which was essential to tackling the problem of paucity of spatial gauged data. This is the method used by the USGS and elsewhere today.

A Federal Interagency Committee undertook an inter-comparison of methods for the prediction of flood frequency discharges at ungauged sites, and the USGS quantile method was found to be minimum variance and least biased of all the models compared (Newton and Herrin 1981). The USGS has through the years made various statistical improvements in their estimation of quantile discharges. Despite all the improvements in the quantile regression method, it has remained a black box statistical method in so far as it does not include any knowledge of physical processes that produce floods from rainfall in space and time. As mentioned in the introduction, due to changes in the hydro-climate system, future outcomes will not be reflected in past hydrologic observations. Therefore, regional predictions of annual flood quantiles will necessarily have to come from the physical processes producing floods that reflect hydro-climate changes rather than from historical observations. Can the results in the USGS regional quantile regression analyses be understood in terms of physical mechanisms producing floods?

As a first step towards achieving this goal, motivated by the ideas of multiscaling in spatial rainfall and river flow distributions (Gupta and Waymire 1990), Gupta, Mesa, and Dawdy (1994) defined regions to be homogeneous if quantile regressions only depend on drainage areas and no other regression variables. Drainage area is the most important, and many times the only variable used; see for example, Smith (1992), Cathcart (2001), and Eaton et al. (2002). Gupta et al. (1995) found that the USGS regional flood frequency relations between annual flood quantiles and drainage areas contain a robust statistical signature as power laws**.** They also attempted to give a heuristic physical interpretation of scaling in annual flood quantiles. However, Gupta et al. (1996) soon after demonstrated how spatial power laws in floods arise at the time scales of RF-RO events with the implication that an extension to annual time scale can be achieved by considering multiple events in a year.

Gupta et al. (1996) considered a Peano basin, which shares the property of mean self-similarity with real networks, as explained in Section 5, but other features of the Peano network are not realistic. They predicted flow hydrographs and resulting peak flows at multiple spatial locations in the interior of a basin for an instantaneous uniform rainfall input and constant velocity of flow in the network. Under these idealized physical conditions pertaining to a RF-RO event, the peak flows were analytically shown to exhibit a spatial power law relationship with respect to drainage area with a flood-scaling exponent, $\phi = \log 3 / \log 4$. The parameter 2ϕ is a *fractal dimension* of the spatial regions of a Peano basin that contribute to peak flows at successively larger drainage areas. Both the power law in floods and the value of the scaling exponent $\log 3 / \log 4$ are a consequence of the mean self-similarity of the Peano basin. This result illustrates a coupling between the geometry of the channel network and physical processes producing floods under highly idealized conditions. Gupta et al. (1996) relaxed the assumption of spatially uniform rainfall and considered an instantaneous spatial rainfall input as a random cascade (see, e.g., Over and Gupta 1994; Over and Gupta 1996; Menabde, Seed, Harris, and Austin 1997). Analytical results showed that random self-similarity in rainfall can be coupled with self-similarity in the topology and geometry of a Peano network, which produces spatial power law statistics in the floods for a RF-RO event. Moreover, rainfall intermittence in a random cascade lowers the exponent ϕ below its value for spatially uniform rainfall.

Gupta et al. (1998a) demonstrated using the well-known random topology model (Shreve 1966, Shreve 1967) that the flood-scaling exponent changes because of finite durations of rainfall events. The scaling relationship for small basins is dominated by rainfall-runoff variability with an exponent, $\phi_1 = 1$, and the scaling relationship for large basins is dominated by network structure and flow dynamics with an exponent, $\phi_2 = 1/2$. As explained in Section 3, this qualitative result holds under more general physical conditions than Gupta et al. (1998a) considered. Gupta et al. (1998a and 1998b) published two self-contained overview papers that described key developments until a decade ago, one focusing on the physical aspects and the other on the mathematical aspects.

A power law, $g(t) = t^{\alpha}$, where α is a constant, arises as a solution of a well known functional equation,

$$g(ts) = g(t)g(s) \tag{1}$$

The functional Eq. (1) is fundamentally connected to the concept of *self-similarity*, which can be either dynamic, or geometric, or statistical. Dynamic similitude laid the foundations for defining scale independent dimensionless numbers, for example, Reynolds, Froude, and Mach numbers. The classical concept of dimensional analysis has been generalized to include *dynamic self-similarity* (Barenblatt 1996). Geometrical self-similarity produces fractal sets (Mandelbrot 1982; Feder 1988), which have found a wide variety of applications, some of which are described in this paper. In hydrologic and many other natural systems, in addition to dynamic and geometric self-similarity, one needs to consider statistical self-similarity due to space-time fluctuations in such systems that cannot be attributed to measurement noise. Statistical self-similarity produces Eq. (1) and a power law appears as its solution in characterizing it (Gupta et al. 1998a).

There exists a body of literature in engineering hydrology that combines statistical rainfall inputs with deterministic RF-RO models (Beven 2003) for understanding the physical origins of flood frequencies at the outlet of a basin (Eagleson 1972; Sivapalan, Wood, and Beven 1990). The focus of this body of work has been at a fixed spatial location in a basin rather than on spatial power law statistics as is the case here. These two goals are fundamentally different. For example, engineering approaches to predicting flood frequencies at the outlet partition a basin into subbasins in an arbitrary manner. The dynamical equations don't change with partitioning (Booij 2005). This assumption disregards the foundational issue of scale described in Klemes (1978) and Dooge (1997). In the physical context of our paper, the challenge of scale arises in specifying physical equations for hillslopes-links on a terrain, and in solving the problem of dynamic parametric complexity that arises in integrating these equations for predicting runoff hydrographs at the bottom of every link, and therefore in all sub-basins of a river basin. Moreover, an arbitrary partitioning of a network into sub-basins destroys the self-similarity of a drainage network that is necessary for predicting spatial power laws in floods from physical processes. Therefore, the existing engineering approaches to flood frequency analysis at the outlet of a basin are not appropriate for understanding dynamical origins of spatial power law statistics in floods.

In summary, research within the past 20 years has shown that the fundamental concepts of geometrical and statistical self-similarity in channel networks provide a common unifying umbrella for understanding the dynamical origins of spatial power law statistics in floods. Our discussion in this paper is focused on key developments since Gupta et al. (1998a and 1998b) published their overview papers.

3 Physical Origins of Statistical Scaling in Floods: Dynamical Formulations and Idealized Case Studies in Peano and Mandelbrot-Vicsek Networks

Results in Gupta et al. (1996) and Gupta et al. (1998a) have been generalized in different directions (Menabde, Veitzer, Gupta, and Sivapalan 2001b; Menabde and Sivapalan 2001a; Veitzer and Gupta 2001; Troutman and Over 2001; Morrison and Smith 2001; Richards-Pecou 2002). We focus on two broad sets of generalizations in this chapter. The first set consists of relaxation of various assumptions on physical processes producing floods but still retaining the assumption of a Peano or a Mandelbrot-Vicsek network. Scaling results have been obtained for annual flood frequencies by extending the dynamics from a single RF-RO event to multiple events at annual time scale; these results are briefly reviewed here. The second set of generalizations consists of analyzing spatial flood scaling relationships in mean and random self-similar networks that describe real river networks. This set of papers is reviewed in Section 5.

3.1 Dynamical formulations governing statistical scaling in floods

The physical origins of spatial scaling statistics in floods are developed from an equation of continuity for a channel network that Gupta et al. (1998a, 1998b) proposed. This equation, along with the challenges that arise in a specification of physical processes at appropriate scales, is explained here. Solving the continuity equation requires a storage-discharge relationship, which is a simple form of momentum equation. Its solutions produce stream flow hydrographs at all junctions in a channel network. They have been used to obtain results on spatial scaling statistics in floods under idealized physical conditions on Peano and Mandelbrot-Vicsek networks, as briefly reviewed in Section 3.2.

Assume that at an appropriate scale of resolution a *drainage network* τ can be represented as a *finite binary rooted tree graph*, consisting of links, junctions and sources, and the outlet. A *junction* is defined as a point where three channels meet. A *source* denotes the starting point of an unbranched channel, and a *link* is a channel segment between two junctions, a source and a junction, or the outlet and a junction. Each link drains two hillslopes on either side of it. The *outlet* is a privileged junction that "directs" or drains water from any location within the entire channel network, and no other junction has this property. It is assumed that a river basin is partitioned into a large number of hillslope-link pairs at the smallest resolvable scale, which can capture the natural partitioning of the terrain of a basin. Mantilla et al. (2005) have demonstrated that the resolution of a digital elevation model (DEM) should be 30 m or less in order to ensure that link-hillslope decomposition of a terrain is not artificially distorted. A decomposition of the terrain of a basin that can capture the natural partitioning into hillslope-link pairs is necessary for developing the scaling theory of floods.

Let $q(e,t)$, $e \in \tau$, $t \geq 0$ be a space-time field representing river discharge, or volume of flow per unit time, through the downstream end of link e. Let $R(e,t)$ denote

the runoff intensity into link e from the two adjacent hillslopes, and let $a(e)$ denote the combined areas of the two hillslopes. $R(e,t)$ may also include the runoff depletion per unit time due to channel infiltration or evaporation from channel surface in link e. Finally, let $S(e,t)$ denote storage in link e at time t. Then the equation of continuity for the system of hillslopes-links can be written as (Gupta et al. 1998a, 1998b),

$$\frac{dS(e,t)}{dt} = -q(e,t) + q(f_1,t) + q(f_2,t) + R(e,t)a(e),\ e \in \tau,\ t \geq 0 \tag{2}$$

The terms $q(f_1,t)$ and $q(f_2,t)$ on the right-hand side represent discharges from the two upstream links. Solutions of Eq. (2) depend on the branching and geometric structure of a channel network through these two terms. An understanding of scaling in peak flows from solutions of Eq. (2) requires some familiarity with self-similarity in channel networks. We give an overview of progress on this important topic in Section 5. Equation (2) as written is based on an assumption that the channel network is binary. It may easily be generalized, however, to non-binary networks, or those for which more than two upstream links can flow into a junction (for example a Peano network); however, this situation does not arise in actual river networks. Equation (2) assumes that no loops are present in the network.

A functional relationship between link storage $S(e,t)$ and link discharge $q(e,t)$ is needed to express Eq. (2) in terms of one dependent variable. It can be obtained from the definition of storage and discharge, and a specification of the link velocity, $v(e,t)$ from a momentum balance equation governing three-dimensional velocity field in an open channel at the sub-link scale. This is a non-trivial problem as Reggiani, Sivapalan, Hassanizadeh, and Gray (2001) pointed out. Kean and Smith (2005) recently have made progress towards solving this problem in the context of predicting a theoretical relationship between flow depth and discharge known as a rating curve. They developed a fluid-mechanical model that resolves boundary roughness elements from field measurements over a natural channel reach and calculated the cross-sectional averaged velocity to predict theoretical rating curve. They also predicted the widely used empirical Manning and Chezy friction parameters, and demonstrated that they are not constant, as is commonly assumed, but vary with flow depth. Kean et al. (2005) work provides a scientific framework to specify $v(e,t)$ for a link.

For simplicity, in this presentation, we follow Menabde et al. (2001a), who assumed a constant Chezy's friction parameter $C(e)$ for a link, and obtained a storage-discharge relationship for a wide channel,

$$S(e,t) = C(e)^{2/3}\gamma(e)^{-1/3}w(e)^{1/3}l(e)q(e,t)^{2/3} \tag{3}$$

Here, $l(e)$ is the link length, $w(e)$ is the link width, and $\gamma(e)$ is the link slope. Substituting Eq. (3) into Eq. (2) gives,

$$\frac{dq(e,t)}{dt} = K(q)\{-q(e,t) + q(f_1,t) + q(f_2,t) + R(e,t)a(e)\}, \tag{4}$$

where, $K(q) = (3/2l)C^{2/3}\gamma^{2/3}w^{-1/3}q^{1/3}$.

A specification of $R(e,t)$ for a hillslope requires a hillslope-scale model for representing runoff generation from Hortonian and saturated overland flows and sub-surface storm flow. Duffy (1996) published a key paper on a hillslope-scale dynamical model to parameterize saturated overland flow and sub-surface storm flow from a detailed three-dimensional representation of unsaturated-saturated flow continuum equations. However, he did not consider Hortonian overland flow, which requires an expression for spatially integrated infiltration on a hillslope in terms of soil moisture, rainfall duration and intensity. It is well known that hydraulic conductivity can vary spatially by orders of magnitude at the scale of tens of meters in a hillslope and that a typical hillslope is anisotropic due to the layered structure of soil strata. Corradini, Govindaraju, and Morbidelli (2002) have recently made progress in parameterizing area-average infiltration on a hillslope in terms of main characteristics of rainfall and soil. However, parametric models of all the three modes of runoff generation comprising $R(e,t)$ are needed in terms of soil-moisture and rainfall characteristics from three-dimensional representation of unsaturated-saturated flow continuum equations. This is an important problem for future research in hillslope hydrology.

Once all the physical processes are specified, Eq. (4) can be solved iteratively to obtain $q(e,t), e \in \tau, t \geq 0$, which represents runoff hydrograph at the bottom of every link, and therefore in all sub-basins of a river basin. Dynamic simulations of flow hydrographs on a channel network involve a large number of hillslope-link pairs. The functions $K(q(e,t))$ and $R(e,t)$ vary from one link-hillslope pair to another. For example, it is well known that the hydraulic-geometric (HG) properties of channels (width, depth, velocity, slope, friction, sediment transport) in a network vary spatially (Leopold, Wolman, and Miller 1964; Ibbit, McKerchar, and Duncan 1998). A specification of HG properties is required for solving Eq. (4); examples of such a specification are given below and again in Section 4. Similarly, soil-moisture and "effective hydraulic conductivity" parameters governing runoff generation vary from one hillslope to another due to spatial heterogeneity among hillslopes. As illustrated in Section 4, the number of parameters increases with the area (scale) of a river basin, and most of these parameters are not measurable. Gupta (2004) calls it the grand challenge of *dynamic parametric complexity*.

Simulated hydrographs $q(e,t), e \in \tau,\ t \geq 0$ at the bottom of every link can be used to compute peak flows for predicting spatial statistical power laws in floods for RF-RO events, and to test these predictions against data. Such comparisons provide the scientific basis for developing a diagnostic framework in real basins and to address the challenge of dynamic parametric complexity, which is explained in Section 4. Clearly, the above problems pertain to the fundamental "challenge of scale" in hydrologic sciences that Klemes (1983) and Dooge (1997) have articulated in a broad context.

3.2 Dynamical origins of scaling in floods in Peano and Mandelbrot-Vicsek networks

Not much is known about general solutions of Eq. (4) within a space-time context, because this line of research is quite new. However, particular cases of analytical and numerical solutions have been obtained for idealized channel network models under idealized physical assumptions. For example, Menabde et al. (2001b) solved Eq. (4) iteratively on a Peano network in space and time in the presence of dynamic storage, $dS(e,t)/dt > 0$, and a constant velocity, and computed flow hydrographs at the end of every link for an instantaneous, uniform precipitation input. It was found that scaling in peak flows holds asymptotically for large drainage areas, but the large-area scaling exponent, $\phi_2 = 0.63$ was found to be below the value, $\log 3/\log 4 = 0.792$ it takes when, $dS(e,t)/dt = 0$, and velocity is a constant (Gupta et al. 1996). Analytical demonstration of this result remains open.

Menabde et al. (2001a) conducted a systematic analysis of the effect of finite duration of rainfall producing flood events, of nonlinear flow dynamics and HG in a network, on the flood scaling exponents ϕ_1 and ϕ_2 on a mean self-similar Mandelbrot-Vicsek network. They found from solving Eq. (4) that the resulting peak flows have an asymptotic power law dependence on drainage area. However, a nonlinear velocity and dynamic storage, $dS(e,t)/dt > 0$, lower the scaling exponent below its value in the presence of dynamic storage and a constant velocity (Menabde et. al. 2001a).

Menabde et al. (2001a) further extended the event-based scaling analysis to annual flood frequencies on the Mandelbrot-Vicsek network using multiple rainfall events and continuous-space-time simulations. They found that spatial scaling between annual quantiles $q(A,T)$ and drainage area A holds for different return periods T and is given by,

$$q(A,T) = A^{\theta_1}\{\sigma - \mu \log\log[T/(T-1)]\} \tag{5}$$

The small-basin annual quantile scaling exponent θ_1 was between 0.60-0.81 and depended on the value of a physical parameter governing infiltration and runoff generation from rainfall events. However, theoretical connections between the small-basin flood scaling exponents ϕ_1 for individual flood events and the small-basin scaling exponent θ_1 for annual flood quantiles was not adequately explored even for this idealized case study. In addition, they did not investigate a connection between the large-basin scaling exponents ϕ_2 for events and the large-basin scaling exponent θ_2 for annual flood quantiles. Nevertheless, above studies provided many key theoretical insights into a connection between hillslope-link scale *physical processes and statistical flood scaling parameters for events and annual flood frequencies*. These studies were a necessary step before going to real basins involving real data as described below.

4 Physical Origins of Statistical Scaling in Floods in Real Basins: The Grand Challenge of Scale and a Diagnostic Framework

In the last five years, the main focus of our research has been to generalize the scaling theory of floods from idealized to real basins that include observed rainfall and stream flows. Many challenges that revolve around the issue of scale arise in this context, and progress has been made on some of them. A central challenge that is explained here is to establish a diagnostic framework to solve the problem of dynamic parametric complexity (Gupta 2004). Such a framework would allow us to test and falsify our physical assumptions by comparing physical predictions of statistical scaling in floods against observations, and to propose new assumptions. The entire issue of diagnosing mathematical models using observations has been and continues to be a central challenge for hydrologic sciences. The goal of diagnosing a mathematical model is quite different from fitting mathematical models to data that has been well established in hydrologic engineering; see Klemes (1997) for a clear explanation of this foundational conceptual and scientific issue. The scaling theory of floods cannot be generalized to real basins for making predictions without a suitable diagnostic framework. Such a framework is the focus of this section.

Progress has been made in producing observational evidence for the existence of power laws in annual flood quantiles as well as in individual RF-RO events, and how the two are related. This body of research has been conducted in two experimental basins of the Agriculture Research Service (ARS), the Walnut Gulch basin in Arizona (Goodrich, Lane, Shillito, and Miller 1997), and the Goodwin Creek experimental watershed (GCEW) in Mississippi (Ogden and Dawdy 2003). Unfortunately, the availability of comprehensive hydrologic data sets that are required to diagnose the scaling theory of floods is rather limited in the US and other parts of the world. The need for new data to test the theory poses a major observational challenge for the future development of the theory that is briefly discussed in Section 6.

4.1 Observational evidence for scaling in annual flood quantiles

Tests of power laws in annual flood quantiles and estimation of corresponding scaling exponents θ_1 and θ_2 require stream flow observations over many years. For example, empirical estimates of 100-year flood quantiles require at least 100 years of data, which is seldom available even for the gauged streams of the world (National Research Council 1988, p. 2). Hydrologists have tried to circumvent this problem by fitting pre-selected flood frequency distributions to annual flood data. However, any such procedure cannot adequately represent the tails of the probability distributions, an issue that remains unsolved. The assumption of stationarity is another critical issue that is discussed in Klemes (1989) and in Section 1.

Goodrich et al. (1997) plotted 2-year and 100-year flood quantiles against drainage areas using annual peak flow data from the eighteen nested subbasins of the Walnut Gulch basin as shown in Fig. 8.1. Log-log linearity between flood quantiles and drainage areas with approximately the same slopes for 2-year and 100-year floods is a signature of *statistical simple scaling* in this data set. We refer to Gupta et al.

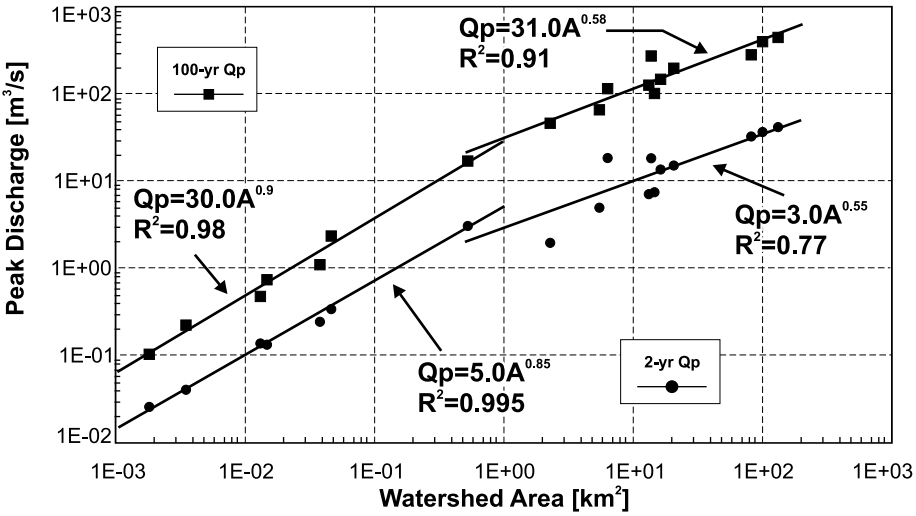


Fig. 8.1. A plot of annual flood quantiles for 2- and 100-year return periods versus drainage areas in the Walnut Gulch basin (Goodrich et al. 1997).

(1998a) for a self-contained explanation of statistical simple scaling. Three features in Fig. 8.1 should be noted: (i) a power law or scaling relationship between peak flows and drainage areas with an exponent $\theta_1 \in (0.85, 0.9)$, which is close to 1 for basins smaller than 1 km^2; (ii) a change in scaling exponents around 1 km^2 (unfortunately, missing gauged observations around 1 km^2 don't allow us to say precisely whether the transition is gradual or abrupt); and (iii) a power law exponent $\theta_2 \in (0.55, 0.58)$, which is close to 0.5 for basins larger than 1 km^2. Goodrich et al. (1997) don't give error bounds for the exponent values.

Ogden et al. (2003) also found a power-law relationship between the 2-year and 20-year annual flood quantiles and drainage areas on the GCEW with a slope, $\theta_1 = 0.77 \pm 0.04$. This result is similar to the finding of Goodrich et al. (1997), and it corresponds to the left hand side of the scaling relationship shown in Fig. 8.1. A change in scaling exponents in GCEW is not apparent from data. We conjecture that a change occurs towards the outlet of the basin, which is 21.2 km^2. The small-to-large basin transition zone in GCEW is therefore apparently much larger than the 1 km^2 observed in the Walnut Gulch basin. These two sets of analyses suggest that the area at which scaling relationship changes depends on space-time rainfall variability, on the nature of the processes of runoff generation on hillslopes and on the runoff dynamics in the two basins. Further insights into this and other important issues require the development of a diagnostic framework as explained below.

In ungauged basins, or in the context of climate change due to global warming, long-term stream flow data sets will not be available (by definition) to conduct quantile regression analyses for predicting annual flood frequencies. The only feasible approach is to first consider an event-based analysis and then develop a predictive understanding of power laws from physical processes by considering multiple events

at the annual time scale. In addition, an event based understanding of flood scaling can be expected to open up new possibilities to improve flood forecasting methods in real time in ungauged basins that are based in RF-RO processes.

4.2 Observational evidence for scaling in floods for rainfall-runoff events and its connection with annual flood quantiles

Is there any observational evidence for spatial scaling in floods for RF-RO events in the Walnut Gulch and GCEW? Ogden et al. (2003) carried out the first observational study pertaining to scaling in flood events. They observed that plots of peak flows against drainage areas for individual RF-RO events on the subbasins of GCEW show a power-law behavior. The regression slope of this relationship can be interpreted as the spatial mean flood-scaling exponent ϕ_1 for an event. Among the 223 events that were analyzed, ϕ_1 was found to vary from 0.9 to 0.6. Fig. 8.2 shows scaling in peak flows for four events from GCEW. One may notice the scaling exponent ϕ_1 and the intercept α_1 vary amongst events. Ogden et al. (2003) computed a mean scaling exponent for the 223 events, $\overline{\phi}_1 = 0.826 \pm 0.047$. Note that, $\theta_1 = 0.77 \pm 0.04 \approx \overline{\phi}_1$. This result gave the first major observational support that the mean $\overline{\phi}_1$ of the event-based scaling exponents and the annual flood quantile-scaling exponent θ_1 are closely related. These two exponents have different definitions, but they appear to be numerically close to each other. *A close relationship between the two exponents suggests a very important hypothesis that it would be possible to predict flood scaling at annual time scales from scaling in individual RF-RO events.*

We analyzed space-time patterns of rainfall events for the Walnut Gulch basin, which has a dense network of raingauges, and identified three sets of populations: 1) events that cover the entire basin in different moments of time due to moving storms over the basin; 2) events that only partially cover the basin; and 3) events that cover almost the entire basin at the same time. Fig. 8.3 for two RF-RO events from the third population set shows a log-log plot of observed peak flows at each of the 19 gauging stations against upstream drainage areas that correspond to each gauge. It exhibits three prominent features for both the events: (i) a power law or scaling relationship between peak flows and drainage areas with an exponent $\phi_1 \in (0.84, 0{,}87)$, which is close to 1 for basins smaller than 1 km^2; (ii) a change in scaling exponent around 1 km^2; and (iii) a power law exponent $\phi_2 \in (0.53, 0.63)$, which is close to 0.5 for basins larger than 1 km^2. Gupta et al. (1998a), and subsequently others, predicted these three features for idealized river networks involving idealized rainfall-runoff events as reviewed in Sections 2 and 3. It is significant to note that these three features are also present in annual flood frequencies in the Walnut Gulch basin as shown in Fig. 8.1.

To investigate further the scaling for individual RF-RO events, we isolated six events from the third population set defined above. These events occurred on 9/9/1964, 8/10/1971, 8/18/1971, 9/13/1975, 9/10/1982 and 9/10/1983. We calculated the mean scaling exponents $\overline{\phi}_1$ and $\overline{\phi}_2$ from these events. Results are as follows:

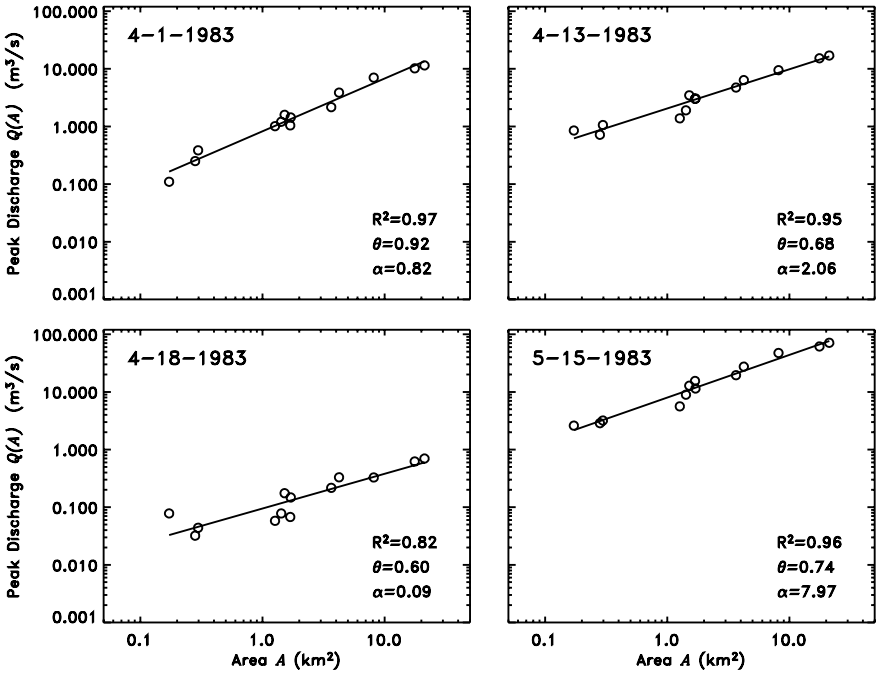


Fig. 8.2. Observed power-laws in peak flows for four rainfall-runoff events in Goodwin Creek basin, MS (Furey and Gupta 2005).

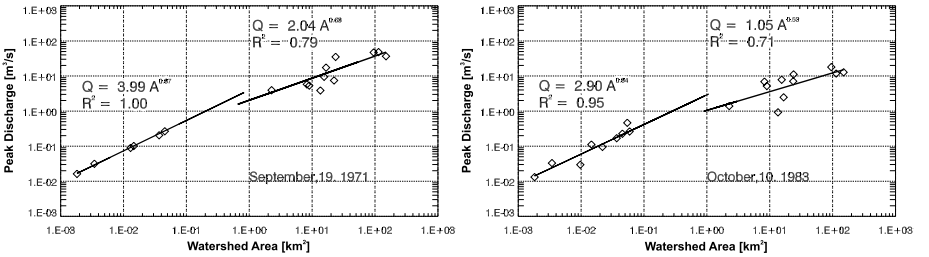


Fig. 8.3. Observed scaling in peak flows for two events on September 19th 1971 and October 10th 1983 in the Walnut Gulch basin.

$\overline{\phi}_1 = 0.77 \pm 0.14$. It can be compared to the exponent $\theta_1 \in (0.85, 0.9)$ from Goodrich et al. (1997) shown in Fig. 8.1. Likewise, $\overline{\phi}_2 = 0.63 \pm 0.11$ compares well with the exponent $\theta_2 \in (0.55, 0.58)$ in Fig. 8.1. This preliminary analysis suggests that for events of the third type there is good agreement between event scaling exponents and scaling exponents observed in the annual flood quantile analysis. However, statistical errors in the estimated exponents from only six events are rather large due to small sample size. By comparison, the statistical errors in Ogden et al. (2003) are much

smaller because they analyzed 223 events, and obtained $\overline{\phi}_1 = 0.826 \pm 0.047$ and $\theta_1 = 0.77 \pm 0.04$. Understanding a precise connection between the scaling of individual RF-RO events and scaling of the maximum annual floods is a very important problem for further research.

4.3 Structure of a diagnostic framework for testing physical assumptions using scaling in floods for rainfall-runoff events

Mantilla et al. (2005) developed CUENCAS, a state-of-the-art, numerical simulation framework for digital drainage basins that is coupled to a GIS. It enables a dynamical integration of the mass balance equation for computing flow hydrographs at the bottom of each link in a channel network. A dynamic integration of Eq. (4) involves a large number of hillslope-link pairs, and a large number of dynamic parameters governing runoff generation and transport. For example, Gupta (2004) has illustrated using an idealized example from Menabde et al. (2001a) that the number of values of physical parameters N at the hillslope-link scale is given by, N=110 A, where A is the drainage basin area. For example, using this estimate in a 21 km^2 basin like GCEW, N=2310. In the Walnut Gulch basin with an area of 150 km^2, N=16,500. Most of these parameters cannot be measured. Therefore, parametric assumptions at the hillslope-link scale cannot be tested directly against observations. This example illustrates the problem of dynamic parametric complexity (Gupta 2004) that was mentioned in Section 3.

For diagnostic purposes, one needs to start with the simplest possible assumptions on hillslope-link scale dynamical parameters, solve Eq. (4) under this set of assumptions, and predict flood-scaling parameters. A dynamical prediction of the scaling parameters under a given set of assumptions allows us to test the discrepancy between theoretical predictions and empirically estimated values of the scaling parameters. Thereby, this approach provides a rigorous theoretical framework to test different physical assumptions and to formulate and test new assumptions. We will illustrate this idea using two new studies that have been conducted on Walnut Gulch and GCEW.

Our first example is taken from Mantilla, Gupta, and Mesa (2006). They conducted a preliminary analysis for an instantaneous and spatially uniform injection of runoff to investigate how three different assumptions on link dynamics affect the prediction of scaling in peak flows on the Walnut Gulch basin. The assumptions were, 1) constant velocity, 2) constant Chezy friction, C, and 3) spatially variable friction that was inversely related to slope $C \sim \gamma^{-\eta}$, which has independent empirical support. Mantilla et. al. (2006) found that power-law scaling in peak flows does not hold for the constant friction scenario. The predicted scaling exponents of peak flows for the other two cases, $\phi_2 = 0.53$ and $\phi_2 = 0.55$, are larger than the scaling exponent of the peaks of the width functions, $\beta = 0.485$. This finding is qualitatively supported by the observed values of the exponent $\phi_2 \in (0.53, 0.63)$ for the two events shown in Fig. 8.3. The theoretical prediction $\phi_2 > \beta$ for a real network is opposite to the previous finding $\phi_2 < \beta$ for the Mandelbrot-Vicsek network that was explained in Section 3 (Menabde et al. 2001a). It illustrates how a diagnostic

framework provides important insights into the dynamics of flow, and spatial variability in physical parameters, in real networks.

The second example is taken from a new study on GCEW that Furey et al. (2005) have conducted for RF-RO events. They analyzed 148 selected events, and observed that both the flood-scaling parameter ϕ_1 and the intercept α_1 are related to the duration of effective rainfall σ. Subtracting the observed stream flow volume from rainfall volume, which gives gross infiltration volume, and estimating a net infiltration parameter that satisfies mass balance was used to estimate σ. Fig. 8.4 shows that the mean scaling exponent $\overline{\phi}_1$ increases with mean duration of effective rainfall $\overline{\sigma}$. Similarly, the intercept $\overline{\alpha}_1$ decreases with respect to an increase of $\overline{\sigma}$. However, the pattern is not as clear for $\overline{\alpha}_1$ due to a lack of data and it is not shown here.

A diagnostic analysis for GCEW has not been conducted. However, it will involve solving Eq. (4) under simplest possible assumptions on hillslope-link scale dynamical parameters for a given RF-RO event, predicting flood hydrographs for every link in a network, and computing the scaling exponent ϕ_1 and the intercept α_1. In addition, the predicted volumes of runoff from hydrographs at all spatial locations where data exist can be compared with the observed to test various assumptions on soil-moisture parameterizations that are required in modeling infiltration and runoff generation on hillslopes as explained in Section 3.1. This type of an investigation is also necessary for diagnosing effective rainfall duration σ for each event. Extensions to

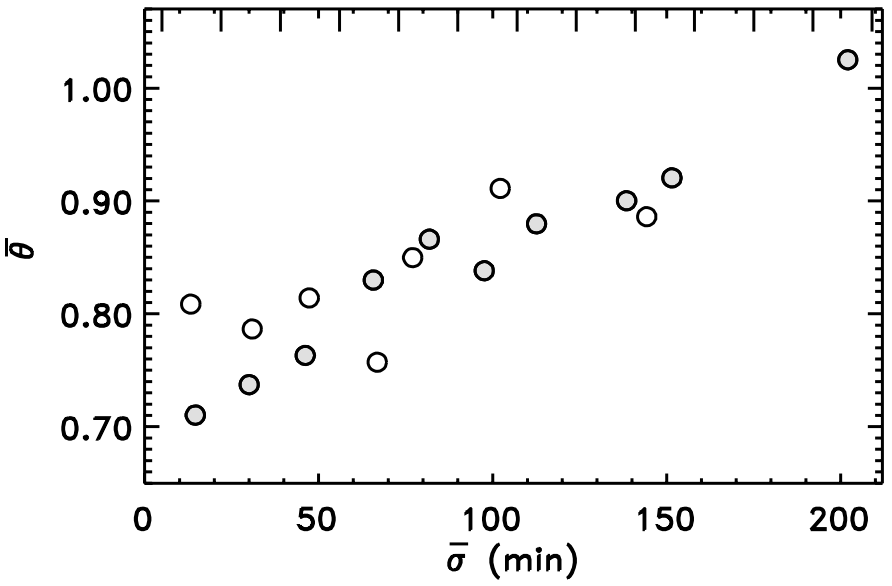


Fig. 8.4. An increase of mean scaling exponents $\overline{\phi}_1$ for rainfall-runoff events with respect to mean durations of effective rainfall $\overline{\sigma}$. Filled and open circles exponent values represent two different seasons. (Furey et al. (2005) used the symbols $\overline{\theta}$ instead of $\overline{\phi}_1$.).

multiple events will be required to diagnose how well the predicted values of $\overline{\phi}_1$, $\overline{\alpha}_1$ and $\overline{\sigma}$ compare with observations in Furey et al. (2005).

To summarize the logic behind a diagnostic framework, we consider two different approaches to hydrologic modeling that Klemes (1983) described in the proceedings of the first workshop on 'scale problems in hydrology': a downward approach and an upward approach. The scaling theory is developing a synthesis of a downward approach (involving the values of scaling exponents and coefficients) and an upward approach involving a predictive physical understanding of the scaling parameters in terms of physical processes at the hillslope-link scale on a terrain. The scaling theory for floods can be formally compared to how observable macroscopic thermodynamic parameters are predicted from unobservable microscopic statistical dynamical equations governing molecular motions (Gupta and Waymire 1983). The flood problem is formally similar to many other multi-scale problems that arise in physical, biological, and social sciences (Turcotte et al. 2002).

5 Self-Similarity in Real Channel Networks and its Applications to Floods: 20 Years of Progress

Power laws or scaling relations are ubiquitous in the quantitative analysis of channels networks and river basin geometry. Empirical studies of river basins for years have shown that relationships between many pairs of variables take the form of a power law. More recently, frequency distributions of variables sampled spatially over a basin have also been shown to exhibit power-law tail behavior, which are a consequence of self-similarity in channel network topology and geometry. Much progress has been made in the last 20 years on two fronts. First, detailed empirical analysis of river basin geometry has been greatly aided by the availability of fine resolution DEMs, allowing power-law exponents to be estimated with considerable accuracy and for detailed analysis of large drainage basins to be carried out. Secondly, progress has been made in understanding how observed power-law behavior in floods arises from self-similar network structure.

A complete understanding of why channel networks exhibit self-similarity must ultimately depend on an understanding of the processes controlling fluvial landscape evolution. This question has been addressed in a large body of literature in recent years, especially in the context of understanding the origin of scale invariance and power-law behavior in three-dimensional landscapes. Much of this work is based on attempts to formulate and solve the partial differential equations governing erosion and fluvial processes in the presence of random forcing. Although a detailed treatment of this question is beyond the scope of our chapter, the reader is referred to Rodriguez-Iturbe and Rinaldo (1997) and Birnir, Smith, and Merchant (2001) for an overview.

In Section 5.1, we summarize a few of the well known empirical relationships. In Section 5.2, we discuss some recently developed self-similar network models that can explain the empirical relations given in Section 5.1. These two sections focus on the developments that have mostly taken place since Rodriguez-Iturbe et al. (1997) published a book on fractal channel networks, and Troutman and Karlinger

(1998) published a review of spatial channel network models. Progress on applications of self-similar channel network models to scaling in floods is reviewed in Section 5.3. It serves as the basis to formulate a *key hypothesis that the physical origins of power laws in floods lie in the self-similarity (self-affinity) of channel networks.*

5.1 Generalized Horton laws and power laws in channel networks

Horton's laws have proven to be fundamental in characterizing self-similar behavior in channel networks, and it may be argued that many of the observed power laws in river basins derive from the self-similar network behavior that is revealed under Horton-Strahler ordering. Under this ordering, when two stream segments of the same order merge, stream order of the out-flowing stream increases by one, and otherwise the order of the out-flowing stream is the higher of the two. The three most widely studied Horton's laws involve N_ω, the number of Strahler streams of order ω, $\overline{L}_\omega$, the average length of streams of order ω, and $\overline{A}_\omega$, the average upstream area of streams of order ω, all of them defined for a channel network. It has been observed that ratios of these quantities in successive orders tend to be independent of order, that is, $N_\omega / N_{\omega+1} \approx R_B$, $\overline{L}_{\omega+1} / \overline{L}_\omega \approx R_L$, and $\overline{A}_{\omega+1} / \overline{A}_\omega \approx R_A$, which are known as classical Horton laws.

Peckham and Gupta (1999) generalized the classical Horton laws for mean drainage areas and mean channel lengths to full probability distributions. Specifically, they gave observational and some theoretical arguments for the well-known random model (Shreve 1967) to show that probability distributions of variables rescaled by their means $X_\omega / \overline{X}_\omega$ collapse into a common probability distribution of some random variable, Z, that is independent of order. It is called distributional simple scaling. After coupling it with the classical Horton law for the means with a Horton ratio R_X, it can be written as,

$$X_\omega \overset{d}{=} (R_X)^{\omega-1} Z, \quad \omega = 1,2,... \tag{6}$$

The symbol $\overset{d}{=}$ denotes equality in probability distributions of random variables on either side of it. The relation in Eq. (6) is known as a *generalized Horton law*. Thus, not only the mean value, but any finite moment, if it exists, obeys Horton laws. As an example, Fig. 8.5 shows the classical and the generalized Horton laws for drainage areas A_ω in the Whitewater basin in Kansas (Mantilla et al. 2005). It represents *statistical self-similarity in ordered networks* (SSSON) of the variable under consideration. Other results along this line are given in Dodds and Rothman (2000).

Among the most widely studied of the observed power-law relationships is $L \approx aA^{\varsigma}$, where L is mainstream length and A is drainage area, and where ς is

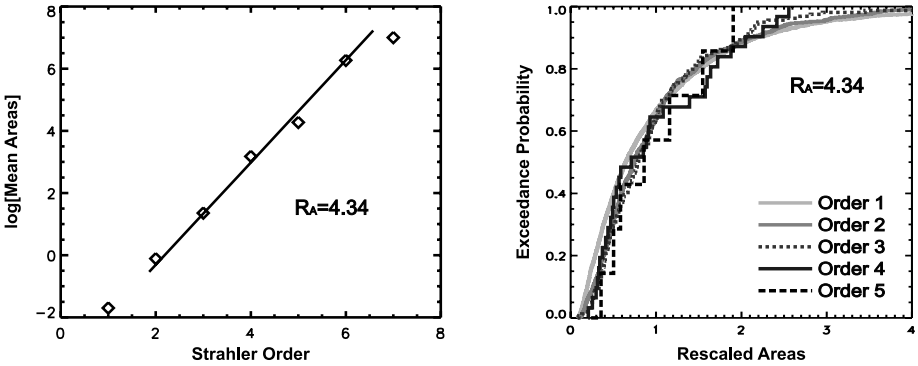


Fig. 8.5. (Left) Horton Law of Mean Areas; (Right) Probability distributions of rescaled areas in the Whitewater Basin, KS (Mantilla et al. 2005).

typically in the range (0.56, 0.6), greater than the value of 0.5 which is predicted by dimensional considerations alone. This relationship has come to be known as Hack's law (Hack 1957). Other power law relationships describe not only the channel network but also the way the network is embedded in a two-dimensional space. For example, if R denotes basin diameter (the maximal distance from the outlet to any point on the boundary), then L tends to scale as R to a power between 1.0 and 1.2, and R scales as A to a power between 0.50 and 0.60. The fact that the scaling exponent in the R vs. A relationship is larger than 0.50 is an indication that larger drainage basins tend to be more elongated than smaller basins, which also partially accounts for the fact that ς is larger than 0.50 (see Rodriguez-Iturbe et al. 1997; Troutman et al. 1998, and Troutman 2005).

The frequency distributions of certain variables sampled spatially over a basin have also been found to exhibit power-law tail behavior. For example, if a frequency analysis of upstream drainage areas A for all channel pixels in a network is performed (extracted from a DEM), the resulting probability of a larger value behaves as a power law, $P(A > a) \sim a^{-\mu}$. The value of the exponent μ is typically observed to lie in the interval (0.43, 0.45) (Rodriguez-Iturbe, Ijjasz-Vasquez, Bras, and Tarboton 1992; Rigon, Rinaldo, Rodriguez-Iturbe, and Ijjasz-Vasquez 1993), and seems to hold over a range of climatic and geologic settings. Histograms of stream lengths similarly exhibit power-law tail behavior.

5.2 Self-similar models of real networks

The random topology model was introduced forty years ago (Shreve 1967), and it has been extensively studied since then. Theoretical prediction of the bifurcation ratio R_B by the random topology model is exactly 4. It holds in the limit as network magnitude $M \to \infty$, where magnitude is defined as the number of first order streams. It has long been assumed that actual networks also have bifurcation ratios that tend to be in the vicinity of 4. Careful recent analyses of data for large networks

(Peckham 1995), however, have revealed that there are in fact significant deviations in actual river networks from the value of 4. Moreover, random model predicts the Hack exponent $\varsigma = 0.50$ (Mesa and Gupta 1987), and $\mu = 0.50$ (DeVries, Becker, and Eckhardt 1994), which do not agree with observed values of these exponents as explained above. Data for large networks have revealed that deviations between random topology model predictions and observations are not due to small sample sizes that give rise to statistical fluctuations around a mean value. The random topology model is clearly inadequate to account for scaling behavior in real networks. These and other results given below instigated the development of the random self-similar model of river networks (Veitzer and Gupta 2000).

Tokunaga (1966) introduced a deterministic model of river networks that was based on Horton-Strahler ordering rather than on link magnitude as is the case for the random topology model. It was characterized by the mean number of side tributaries. For a stream of order ω, let $T_{\omega k}$ denote average number of side tributaries of order *k*. Tokunaga (1966, 1978, 1984) introduced $T_{\omega k}$, known as generators. Generators are self-similar if they obey the constraint, $T_{\omega,\omega-k} = T_k$ independent of ω. Topologically self-similar networks (SSN) are constructed using self-similar generators. Under the additional constraint that $T_{k+1}/T_k \approx R_T$ is a constant, generators have the form, $T_k = T_1 R_T^{k-1}$ (Tokunaga 1966).

It is easy to show that, $N_k, k = 1,2,\ldots$, the number of streams of different Horton-Strahler order, obey a recursion equation (Tokunaga 1978),

$$N_k = 2N_{k+1} + \sum_{j=1}^{\Omega-k} T_j N_{k+j} = 2N_{k+1} + T_1 \sum_{j=1}^{\Omega-k} R_T^{j-1} N_{k+j} \tag{7}$$

Peckham (1995) considered the asymptotic behavior, in the limit of large network order Ω, of several quantities that are a function of Tokunaga generators, including stream numbers N_ω, links per stream C_ω, and magnitude M_ω. He gave a heuristic argument to show that these quantities obey Horton's laws with ratios R_B, R_C and R_M, and that the ratios are related to generator properties. He also showed that the Hack exponent $\varsigma = \log(R_C)/\log(R_B)$. In a later work, Peckham and Gupta (1999) showed that the tail of the cumulative magnitude distribution for Tokunaga networks exhibit the form of a power law with scaling exponent $1-\varsigma$, revealing the important connection between these two scaling exponents. The tail behavior of the magnitude distribution is close to that of the upstream drainage area distribution, which implies that the area probability exponent and the Hack exponent are related as,

$$\mu + \varsigma = 1 \tag{8}$$

Note that the sum of the observed values of the Hack exponent (0.56, 0.6) and the area distribution exponent (0.43, 0.45) given in Section 5.1 is near 1.

Tokunaga also showed that the generators of the random topology model exhibit Tokunaga self-similarity with parameters $T_1 = 1, R_T = 2$, which predicts $R_B = 4$. McConnell and Gupta (2007) have rigorously proved Horton law of stream numbers as asymptotic results from solutions of recursion Eq. (7), and extended it to a Horton law for magnitude M_ω. Newman, Turcotte, and Gabrielov (1997) discuss space-filling characteristics of deterministic spatially embedded trees that obey the Tokunaga self-similarity condition.

The random model is equivalent to the widely studied critical, binary Galton-Watson branching process. Burd, Waymire, and Winn (2000) have shown that the random model is the only model among finite, binary Galton-Watson stochastic branching trees that exhibits the mean self-similar topology of Tokunaga networks. But, as already mentioned, modern observations from large basins show that the random model predictions deviate substantially from empirical values (Peckham 1995). These findings opened the door to develop a new class of statistical channel network models called the random self-similar networks (RSN) (Veitzer et al. 2000; Troutman 2005). RSNs do not belong to the class of binary Galton-Watson stochastic branching processes.

Recursive replacement algorithms provide a general and flexible approach for generating networks with self-similar characteristics. Although deterministic constructions have been around for some time, Veitzer et al. (2000) introduced the RSN model. It allows for random generators to better account for the spatial variability of network configurations seen in nature, but these generators are not defined in terms of the side branching structure as in Tokunaga networks. Their formulation also allows for different generator probability distributions for replacement of interior and exterior links, where an exterior link is one with no upstream inflows. Veitzer et al. (2000) showed that the distributional simple scaling holds for the number of links per stream, magnitude, and stream numbers as network order $\Omega \to \infty$, and the average side tributary structure of a subset of RSNs obey Tokunaga self-similarity. Troutman (2005) showed how RSN generators may be uniquely extracted from actual networks under certain restrictions on the class of generators and gave results showing that a geometric distribution of the number of side tributaries in both interior and exterior generators fit reasonably well for the Flint River basin in Georgia. Estimates of the geometric parameters were used to show that a version of Hack's law (based only on topological network properties) with exponent 0.62 holds for the Flint River. Mantilla (2007) and Mantilla, Troutman, and Gupta (2007) have done further testing of this model for 28 basins in diverse climatic and geographic settings in the US, again showing good agreement of data with the geometric distribution in all cases.

Veitzer, Troutman, and Gupta (2003) considered power law tail behavior of the upstream area distribution under a different sampling scenario than Rodriguez-Iturbe et al. (1992). Whereas previous studies had been based on a uniform sampling of all links in a basin, Veitzer et al. (2003) looked at the area distribution if only links which drain complete Horton-Strahler networks are sampled. It was shown that for deterministic SSNs the scaling exponent is 1, and further, if areas are randomized in such a way as to preserve statistical simple scaling, the area distribution-scaling exponent under this sampling scenario is $\log(R_B)/\log(R_A)$. Data analyses of tail

probabilities for the Flint River basin show much better power law behavior when only complete Horton-Strahler links are sampled, suggesting that it is the self-similar geometry of the channel network that is the most important factor in defining power laws for spatially sampled variables. Therefore, an effective sampling strategy for accurate estimation of scaling exponents is to sample at outlets of complete Horton-Strahler streams. This conclusion has important implications for designing of sampling networks for looking at scaling in flood peaks. For example, the Hydro-Kansas (HK) interdisciplinary pilot project involving several principal investigators, including the authors of this paper, has located 11 stream gauges at the end of complete Horton-Strahler streams in the 1100 km^2 Whitewater basin in Kansas.

An important consideration in analysis of scaling properties of river basins and the channel networks that drain them is spatial embedding, which is also necessary for simulating spatially distributed rainfall intensity fields on river basins. The idealized basins, Peano and Mandelbrot-Vicsek, considered in Sections 2 and 3 are spatially embedded, which enabled a spatial distribution of rainfall intensities from a random cascade model on these basins. For networks generated by models such as the random topology model, deterministic SSTs, and RSNs, the branching structure is defined but how the networks fit into two-dimensional space is not. One way to model spatially embedded networks is to define them on a lattice, i.e., they are "lattice trees". Often networks in this class of models are constructed using rule variations for self-avoiding random walks on a lattice (Leopold and Langbein 1962). See, Meakin, Feder, and Jossang (1991) for examples of how such rule variations affect power law scaling in networks. Much research has been done on scaling properties of *optimal channel networks* (Rodriguez-Iturbe et al. 1997). These are lattice trees, which minimize a certain network function having an interpretation in terms of energy expenditure. They are a special case of Gibbsian networks that Troutman and Karlinger (1994, 1998) have investigated. Many of the scaling exponents in such models can be obtained only by numerical simulation, as rigorous analytical results are intractable. Mantilla (2007) has recently made progress on the spatial embedding problem where an algorithm is presented for embedding RSNs in a given spatial region. It seems feasible to obtain analytical results for embedded networks using this algorithm. This work will be published in the future.

5.3 Self-similarity in channel networks and its implications for power laws in floods

The fundamental role of network geometry in influencing streamflow response has been thoroughly explored in recent years. It was argued in a number of early works that the network width function is the logical starting point for seeking the connection between channel network properties and streamflow hydrographs (Kirkby 1976; Lee and Delleur 1976; Gupta et al. 1983; Troutman and Karlinger 1984; Gupta and Mesa 1988). For any given distance from the outlet, the width function is defined to be number of streams at that distance. It is seen that if rainfall is deposited instantaneously and uniformly over a network and travels at a constant velocity to the outlet, then the flow hydrograph at the outlet will have the same shape as the width function. The width function is one version of the so-called geomorphologic instantaneous

unit hydrograph (GIUH). Another version of a GIUH that Rodriguez-Iturbe and Valdes (1979) and Gupta, Waymire, and Wang (1980) proposed, and on which a substantial literature exists, is not considered here. For a recent review, see Rinaldo and Rodriguez-Iturbe (1996).

Troutman et al. (2001) generalized the results on scaling exponents of peak flows in Gupta et al. (1996) from a Peano basin to a more general class of networks generated by a deterministic recursive replacement model, such as the mean SS Tokunaga networks. Width functions for such a model exhibit multifractal behavior. Rainfall in this study was assumed, as in Gupta et al. (1996), to obey a multifractal random cascade model, and it was demonstrated how scaling of flows depends on network and rainfall scaling. Other studies have looked in more detail at multifractal modeling of the width function (e.g., Veneziano, Moglen, Furcolo, and Iacobellis 2000) and at implications of such scaling behavior for floods. Richards-Pecou (2002) estimated multifractal width function properties for 16 Australian networks and showed some relation of the width function scaling properties and flood peak properties.

Veitzer et al. (2001) examined statistical self-similarity of width function maxima for three actual river basins in the U.S. and for the RSN model. They showed that for the actual basins a Horton-type law holds for these maxima, $\overline{\Theta}_\omega / \overline{\Theta}_{\omega-1} \approx R_\Theta$, where $\overline{\Theta}_\omega$ is the average maximum width function for order ω streams, and further that distributional simple scaling also holds. They found that these properties also hold for networks generated by the RSN model. These findings are important in their implications for flood scaling, because under ideal assumptions (particularly constant velocity flow and short duration rainfall), the scaling peak flows will be identical to that of the width function maximum. Troutman (2005) looked further at scaling of the width function for the most general class of RSNs, showing that scaling of average distance to the outlet in these networks depends on mean distance between replacement nodes in interior generators, which coincides with the Horton ratio R_C for a certain class of RSNs.

Menabde et al. (2001b) investigated the effect of linear flow routing incorporating attenuation and network topology on peak flow scaling exponent for an instantaneously applied uniform runoff on simulated deterministic and RSNs. The analysis of simulated hydrographs showed that peak flows have an asymptotically scaling dependence on the drainage area for a subclass of RSNs. Moreover, the predicted scaling exponents were shown to be different from that predicted by the scaling properties of the maxima of the width functions (Veitzer et al. 2001).

Mantilla et al. (2007) in a newly completed study, which is in the process of publication, have performed a comprehensive study of flow simulations on randomly generated RSNs under assumptions of instantaneous, spatially uniform rainfall and constant flow velocity in all links. A general analytical expression for the scaling exponent for the maxima of the expected value of the width functions in terms of R_B and R_C is given, and numerical results are given for the scaling exponent of the expected value of the maxima of the width functions. These two scaling exponents are found to differ, with the latter being larger. The two analogous scaling exponents for peak flows were obtained and compared to scaling exponents of the width function maxima. The peak flow exponents were found to be independent of flow velocity

and rainfall intensity. Mantilla (2007) has also obtained numerical results on the role of HG on peak-flow scaling exponents that will be published in the future.

6 A Brief Discussion of Future Challenges

Through an overview of progress in developing a nonlinear geophysical theory of floods in the last 20 years, we have illustrated how the theory synthesizes disparate concepts into a coherent whole, and provides a theoretical framework around the central issue of scale. The reader may ask the question: How can the scaling theory make statistical predictions of floods under climate change? Our contention is that power-law relationships contain the essential signature of watershed response across a range of scales for the following reason. The basic conservation principles that govern the transport and transformation of water do not inherently contain features that lead to scaling in floods. It is only their manifestation on the self-similar structures of a river network that could lead to power laws. This hypothesis is supported by observations and theoretical developments as explained in this chapter. Even in a regime of changing climate, power laws can be expected to hold, because they are intimately tied to the self-similar structure of networks. The key to predicting watershed response and floods under changing climate is thus to understand the relationships between the parameters in power-law relationships and the climatic factors driving them. It includes rainfall and its variability in space and time, and temperature and atmospheric water vapor driving evapotranspiration and antecedent moisture conditions. A necessary step towards this goal is to develop numerical models of watershed response that are able to consistently predict statistical scaling exponents and intercepts for floods from conservation principles under present-day conditions (Mantilla et al. 2006). Once these models have been established, they can be used to examine how scaling parameters in floods are altered by climate change.

Two sets of challenges arise in future developments of the scaling theory of floods in addition to those already covered in the previous sections. The first challenge pertains to understanding the physical origins of clustering and scaling statistics in space-time rainfall, and the second pertains to physical origins of scaling statistics in HG properties of river networks. Both these topics appear prominently in developing the nonlinear geophysical theory of floods. Understanding scaling in rainfall, the first challenge, is necessary to extend the theory to different climates at the scale of the continents and planet Earth. It is also important in linking the changes in the planetary water cycle due to global warming to changes in flood statistics in river basins of the world. We will discuss both these challenges very briefly. A more extended discussion will be given in future publications.

Two sets of developments have taken place in statistical modeling of space-time rainfall intensities at the meso-scale since the 1980s. The first pertains to statistical clustering in rainfall that the point-process theories captured (see Cox and Isham (1998) for a brief overview), and the second to scale invariance, which the random cascade theory captured (see Foufoula-Georgiou (1998) for a review). A long-standing question is to understand the dynamical origins of these statistical features of rainfall in space and time. Nordstrom and Gupta (2003) published a paper as a

first step to a grand unification in rainfall modeling. They constructed a simple nonlinear dynamical model for oceanic convective rainfall without invoking any statistical assumptions. The space-time solutions of this model exhibited statistical scale invariance found in random cascades as an emergent phenomenon. Variation in the model's physical parameters was shown to have a big effect on the statistical structure of rainfall. This work is in its infancy and much more research is needed on this topic. This line of research will be directly applicable to predicting changes in the statistical structure of rainfall due to changes in climate and the global water cycle.

The challenge of understanding scaling in HG of river networks goes back to the 1950s, when Leopold and Miller (1956) found that the classical Horton laws could be extended to the HG variables on river networks. Prior to this work, Leopold and Maddock (1953) had proposed HG relationships at a station that refers to temporal variability, and in down-stream direction that refers to spatial variability for a channel reach (Leopold 1953). Unique sets of HG observations in two New Zealand basins have shown that significant fluctuations around the means are present in data that are well described as power laws with respect to drainage areas (Ibbit et al. 1998; McKerchar, Ibbitt, Brown, and Duncan 1998). Recent developments have taken place in understanding HG in down-stream direction using ideas of hydraulic similitude (Griffith 2003). However, research is needed to develop a dynamical theory of HG in channel networks that can predict the scaling exponents as well as the fluctuations around the means. We conjecture that self-similarity in channel networks would play a key role in understanding the physical origins of statistical scaling in HG of channel networks.

How can the scaling theory of floods be generalized to global basins spanning different climates at the scale of the continents and the entire planet Earth? This is a big challenge for future research. Poveda, Vélez, Mesa, Cuartas, Barco, Mantilla, Mejía, Hoyos, Ramírez, Ceballos, Zuluaga, Arias, Botero, Montoya, Giraldo, and Quevedo (2007) have taken a first step in this direction, developing a new physical approach to regionalization of annual flood frequencies. Their study involved conducting annual water balances over several drainage basins for the entire country of Colombia, which has widely varying climates ranging from extremely humid to arid and semi-arid. It predicted annual streamflows from interpolated fields of annual precipitation and evapotranspiration using ground-based and remotely sensed data, finding reasonable agreement with observed streamflows. In particular, the best estimates of evapotranspiration were given by the well-known Budyko equation (Budyko 1974). Choudhury (1999) has demonstrated that the Budyko equation can very well predict the annual evapotranspiration on a broad range of spatial scales. This demonstration used observations at the plot scale and numerical simulations from semi-empirical physical models in large drainage basins of the world.

Poveda et al. (2007) showed that power laws describe the relationship between annual flood quantiles and drainage areas, and the flood scaling parameters can be expressed as functions of annual runoff obtained from the water balance. A better understanding of connection between scaling in floods for RF-RO events and annual flood quantiles that is reviewed here, and its link with the approach that Poveda et al. (2007) have taken, offers a concrete direction to extend the scaling theory of floods

to global basins and to longer time scales than annual. A link to planetary hydro-climate would come from estimation and prediction of annual precipitation and evapotranspiration from global climate models. In this context, we refer to the recent paper of Koster and Suarez (1999), who have demonstrated that a function derived from the Budyko equation can be used to relate the interannual variability in evapotranspiration to the gross characteristics of the atmosphere based in annual radiation and interannual variability in precipitation. They tested their analyses using GCMs simulations and found good agreement with predictions from the Budyko equation. This line of investigation provides a new research direction for making flood predictions under a changing global hydro-climate due to global warming.

Last but not least, we note that the scaling framework for floods can be generalized to include ecological and biogeochemical processes coupled to water. As an example, consider riparian evapotranspiration (RET). Based on digital satellite images, we have been able to discern scaling in the areas occupied by riparian vegetation in the Whitewater basin in Kansas. It suggests that scaling in RET exists (Bruce Milne, Professor of Biology, University of New Mexico, personal communication). It is part of the HK project that was mentioned previously in Section 5.2. A dynamical understanding of scaling in RET from diurnal to annual and longer time scales, similar to scaling floods, can be developed by extending the hillslope-link water balance to include RET. The HK investigators are currently exploring this important line of research involving land-atmosphere interactions. We hypothesize that similar scaling relationships exist, yet to be uncovered, in biogeochemical fluxes of nitrogen, carbon and other elements and nutrients.

Acknowledgements

We gratefully acknowledge help from Ricardo Mantilla in producing Fig. 8.3 and in conducting a preliminary analysis of scaling exponents for the Walnut Gulch basin in Section 4.2. In addition, we acknowledge several insightful discussions, as part of the HK project, with Ricardo, Peter Furey, Keith Nordstrom, Hari Rajaram, John Mobley, Bruce Milne, Jason Kean, Jordan Clayton, Jim Smith, Bob Grossman, Bill Eichinger, Ben Balsley, and Jim Butler over the past several years, which greatly impacted our thinking. We gratefully acknowledge leadership of Witek Krajewski and Anton Kruger, University of Iowa, in guiding the field program of the HK project. We appreciate and acknowledge the thorough reviews and helpful comments from Vit Klemes, Thomas Over, and Ken Eng, which greatly improved the paper.

We are grateful to the Army Research Office and the National Aeronautics and Space Administration for funding our research in the early stages of development of the scaling theory of floods. We acknowledge timely seed funding from Marshall Moss, US Geological Survey, which served as a catalyst to initiate collaboration between Dave Dawdy and Vijay Gupta on our research. The Hydrologic Science Program, the Graduate Research Traineeship (GRT) Program, and the Water Cycle Initiative, National Science Foundation, have provided crucial support for this research for over a decade. All of this support is gratefully acknowledged, without which the development of the scaling theory of floods would not have been possible.

We gratefully acknowledge the support of the ARM Climate Research Facility (ACRF) of the Department of Energy for the field component of the HK project.

The high-quality hydrologic observations from the USDA-Agricultural Research Service (ARS) Goodwin Creek and Walnut Gulch Experimental Watersheds were critical for the analysis herein. We gratefully acknowledge the vision and commitment of early ARS and Soil Conservation Service scientists and administrators to construct and operate these and the entire ARS National Experimental Watershed Network for the long-term. The dedication of the ARS staff in maintenance of these long-term hydrologic observatories and diligent long-term collection of high quality hydrologic and watershed data is to be commended and is gratefully acknowledged. Without the availability of these data sets, the observational and the diagnostic component of our research would not have been possible.

References

Band, L. (1986) Topographic partition of watersheds with digital elevation models. Water Resour. Res. 22, 15-24.

Barenblatt, G. I. (1996) *Scaling, Self-Similarity and Intermediate Asymptotics*. Cambridge Texts in Applied Mathematics, 14, Cambridge, UK.

Beven, K. (2003) *Rainfall-Runoff Modelling: the Primer*. John Wiley and Sons Ltd.

Birnir, B., Smith, T.R., and Merchant, G.E. (2001) The scaling of fluvial landscapes. Comp. Geosci. 27, 1189-1216.

Booij M. J. (2005) Impact of climate change on river flooding assessed with different spatial model resolutions. J. Hydrol. 303, 176–198.

Budyko, M.I., (1974) *Climate and Life*, (D. H. Miller (Ed.), English Edition), Academic Press, New York.

Burd, G.A., Waymire, E. and Winn, R.D. (2000) A self-similar invariance of critical binary Galton-Watson trees. Bernoulli. 6, 1-21.

Cathcart, J. (2001) The effects of scale and storm severity on the linearity of watershed response revealed through the regional L-moment analysis of annual peak flows, Ph.D. dissertation, University of British Colombia, Vancouver, Canada.

Choudhury, B.J. (1999) Evaluation of an empirical equation for annual evaporation using field observations and results from a biophysical model. J. Hydrol. 216, 99-110.

Corradini, C., Govindaraju, R.S. and Morbidelli, R. (2002) Simplified modeling of areal average infiltration at the hillslope scale. Hydrol. Process. 16, 1757-1770.

Cox, D. R., and Isham, V. (1998) Stochastic spatial-temporal models for rain. In: O. E. Barndorff-Nielsen, V. K. Gupta, V. Perez-Abreu, and E. C. Waymire (Eds.), *Stochastic Methods in Hydrology: Rainfall, landforms and floods*. Adv. Ser. Stat. Sci. Appl. Prob. 7, World Scientific, pp. 1-24.

de Vries, H., Becker, T. and Eckhardt, B. (1994) Power law distribution of discharge in ideal networks. Water Resour. Res. 30, 3541-3544.

Dodds, P.S. and Rothman, D.H. (2000) Geometry of river networks. I. Scaling, fluctuations, and deviations. Phys.Rev. E. 63, 16115-1-16115-13.

Dooge J. C. I. (1997) Scale problems in hydrology. In: N. Buras (Ed.), *Reflections on hydrology Science and Practice,* American Geophysical Union, pp. 85-143.

Duffy, C. J. (1996) A two-state integral-balance model for soil moisture and groundwater dynamics in complex terrain, Water Resour. Res. 32, 2421-2434.

Eagleson, P. S. (1972) Dynamics of flood frequency. Water Resour. Res. 8, 878-898.

Eaton, B., Church, M. and Ham, D. (2002) Scaling and regionalization of flood flows in British Columbia, Canada, Hydrol. Process. 16, 3245-3263.

Feder, J. (1988) *Fractals*. Plenum Press, New York.

Foufoula-Georgiou, E. (1998) On scaling theories of space-time rainfall. Some recent results and open problems. In: O. E. Barndorff-Nielsen, V. K. Gupta, V. Perez-Abreu, and E. C. Waymire (Eds.), *Stochastic Methods in Hydrology: Rainfall, landforms and floods*. Adv. Series Stat. Sci. and Appl. Prob. 7, World Scientific, pp. 25-72.

Furey, P.R. and Gupta, V.K. (2005) Effects of excess rainfall on the temporal variability of observed peak discharge power laws, Adv. Water Resour. 28, 1240-1253.

Goodrich, D.C., Lane, L.J., Shillito, R.M. and Miller, S. (1997) Linearity of basin response as a function of scale in a semiarid watershed. Water Resour. Res. 33, 2951-2965.

Griffiths, G.A. (2003) Downstream hydraulic geometry and hydraulic similitude, Water Resour. Res. 39, doi: 10.1029/2002WR001485.

Gupta, V. K. (2004) Emergence of statistical scaling in floods on channel networks from complex runoff dynamics. Chaos, Solitons and Fractals. 19, 357-365.

Gupta, V. K., Castro, S. and Over, T.M. (1996) On scaling exponents of spatial peak flows from rainfall and river network geometry. J. Hydrol. 187, 81-104.

Gupta, V. K. and Dawdy, D. (1995) Physical interpretation of regional variations in the scaling exponents in flood quantiles. Hydrol. Proc. 9, 347-361.

Gupta, V.K. and Mesa, O.J. (1988) Runoff generation and hydrologic response via channel network geomorphology: Recent progress and open problems. J. Hydrol. 102, 3-28.

Gupta, V. K., Mesa, O.J. and Dawdy, D.R. (1994) Multiscaling theory of flood peaks: Regional quantile analysis. Water Resour. Res. 30, 3405-3421.

Gupta V. K. and Waymire, E. (1983) On the formulation of an analytical approach to hydrologic response and similarity at the basin scale. J. Hydrol. 65, 95-123.

Gupta, V. K. and Waymire, E. (1990) Multiscaling properties of spatial rainfall and river flow distributions. J. Geophys. Res. 95, 1999-2009.

Gupta, V. K. and Waymire, E. (1998a) Spatial variability and scale invariance in hydrologic regionalization. In: G. Sposito (Ed.), *Scale Dependence and Scale Invariance in Hydrology*, Cambridge University Press, pp. 88-135.

Gupta, V.K. and Waymire, E. (1998b) Some mathematical aspects of rainfall, landforms and floods. In: O. E. Barndorff-Nielsen, V. K. Gupta, V. Perez-Abreu, and E. C. Waymire (Eds.), *Stochastic Methods in Hydrology: Rainfall, landforms and floods*. Adv. Series Stat. Sci. and Appl. Prob. 7, World Scientific, pp. 129-171.

Gupta, V.K., Waymire, E. and Wang, C.T. (1980) A representation of an instantaneous unit hydrograph from geomorphology. Water Resour. Res. 16, 855-862.

Hack, J.T. (1957) Studies of longitudinal stream profiles in Virginia and Maryland. *US Geological Survey Professional Paper* 294-B.

Hobbins, M. T., Ramirez, J., and Brown, T.C. (2004) Trends in pan evaporation and actual evapotranspiration in the conterminous U. S.: Paradoxical or complementary? J. Geophy. Res. 31, L13503, doi: 1029/2004GL019846.

Ibbitt, R. P., McKerchar, A.I., and Duncan, M.J. (1998) Taieri river data to test channel network and river basin heterogeneity concepts. Water Resour. Res. 34, 2085-2088.

IPCC (Intergovernmental Panel on Climate Change) (2001) Third Assessment Report, Climate Change 2001, World Meterol. Assoc. and United Nations Environ. Programme, Geneva.

Karl, T. R., Knight, R.W., Easterling, D.R. and Quayle, R.G. (1996) Indices of climate change for the United States. Bull. Am. Meteorol. Soc. 77, 179-292.

Kean, J.W. and Smith, J.D. (2005) Generation and verification of theoretical rating curves in the Whitewater river basin, Kansas. J. Geophy. Res. 110, F04012, doi:10.1029/2004JF000250.

Kirkby, M.J. (1976) Tests of the random network model and its application to basin hydrology. Earth Surf. Proc. Landforms. 1, 97-212.
Klemes, V. (1978) Physically based stochastic hydrological analysis. Adv. Hydroscience. 11, 285-356.
Klemes, V. (1983) Conceptualization and scale in hydrology. J. Hydrol. 65, 1-23.
Klemes, V. (1989) The improbable probabilities of extreme floods and droughts. In: O. Starosolsky and O. M. Meldev (Eds), *Hydrology and Disasters*. James & James, London, pp. 43-51.
Klemes, V. (1997) Of carts and horses in hydrologic modeling. J. Hydrol. Eng. 2, 43-49.
Koster, R. D. and Suarez, M.J. (1999) A simple framework for examining the interannual variability of land surface moisture fluxes. J. Climate. 12, 1911-1917.
Lee, M.T. and Delleur, J.W. (1976) A variable source area model of the rainfall-runoff process based on the watershed stream network. Water Resour. Res. 12, 1029-1035.
Leopold, L.B. and Langbein, W.B. (1962) The concept of entropy in landscape evolution. *US Geological Survey Professional Paper* 500A.
Leopold, L. B., Wolman, M.G. and Miller, J.P. (1964) *Fluvial Processes in Geomorphology*, W. H. Freeman, San Francisco.
Leopold, L. B. and Maddock, T. (1953) The hydraulic geometry of stream channels and some physiographic implications. *US Geological Survey Professional Paper* 252.
Leopold, L. B. (1953) Downstream change of velocity in rivers. Am. Jour. Sci. 251, 606-624.
Leopold, L. B., and Miller, J.P. (1956) Ephemeral Streams: Hydraulic Factors and Their Relation to Drainage Net. *U.S. Geological Survey Professional Paper* 282A.
Manabe, S. (1997) Early development in the study of greenhouse warming: The emergence of climate models. Ambio. 26, 47-51.
Mandelbrot, B. (1982) *The Fractal Geometry of Nature*. Freeman, San Francisco, USA.
Mantilla, R. (2007) Physical basis of statistical self-similarity in peak flows in random self-similar networks, PhD dissertation, University of Colorado, Boulder.
Mantilla, R. and Gupta, V.K. (2005) A GIS numerical framework to study the process basis of scaling statistics on river networks. IEEE Geophysical and Remote Sensing Letters. 2, 404-408.
Mantilla, R., Gupta, V.K. and Mesa, O.J. (2006) Role of coupled flow dynamics and real network structures on Hortonian scaling of peak flows. J. Hydrol. 322, 155-167.
Mantilla, R., Troutman, B.M. and Gupta, V.K. (2007) Statistical scaling of peak flows and hydrograph properties in random self-similar river networks: constant velocity case (preprint).
McConnell, M. and Gupta, V.K. (2007) A proof of the Horton law of stream numbers for the Tokunaga model of river networks. Fractals (in press).
McKerchar, A.I., Ibbitt, R.P., Brown, S.L.R. and Duncan, M.J. (1998) Data for Ashley river to test channel network and river basin heterogeneity concepts. Water Resour. Res. 34, 139-142.
Meakin, P., Feder, J. and Jossang, T. (1991) Simple statistical models for river networks. Physica A. 176, 409-429.
Menabde, M., Seed, A., Harris, D. and Austin, G. (1997) Self-similar random fields and rainfall simulation, J. Geophys. Res. 102, 13509-13515.
Menabde, M. and Sivapalan, M. (2001a) Linking space-time variability of river runoff and rainfall fields: a dynamic approach. Adv. Water Resour. 24, 1001-1014.
Menabde, M., Veitzer, S.E., Gupta, V.K. and Sivapalan, M. (2001b) Tests of peak flow scaling in simulated self-similar river networks. Adv. Water Resour. 24, 991-999.
Mesa, O.J. and Gupta, V.K. (1987) On the main channel length-area relationship for channel networks. Water Resour. Res. 23, 2119-2122.
Morrison, J. A., and Smith, J. (2001) Scaling properties of flood peaks. Extremes. 4, 5-22.
National Research Council (1988) Estimating Probabilities of Extreme Floods: Methods and Recommended Research. National Academy Press, Washington, D.C.

National Research Council (1991) *Opportunities in the Hydrologic Sciences.* National Academy Press, Washington D.C.

Newman, W.I., Turcotte, D.L. and Gabrielov, A.M. (1997) Fractal trees with side branching. Fractals. 5, 603-614.

Newton, D. W. and Herrin, J. C. (1981) Assessment of Commonly Used Flood Frequency Methods, paper presented at Fall Meeting, Amer. Geophys. Union, San Francisco, CA.

Nordstrom, K. and Gupta, V.K. (2003) Scaling statistics in a critical, nonlinear physical model of tropical oceanic rainfall. Nonlinear Proc. Geophy. 10, 531-543.

Nordstrom, K., Gupta, V.K. and Chase, T. (2005) Role of the hydrological cycle in regulating the climate of a simple dynamic area fraction model, Nonlinear Proc. Geophy. 12, 741-753.

Ogden F. L. and Dawdy, D.R. (2003) Peak discharge scaling in small Hortonian watershed. J. Hydrol. Engr. 8, 64-73.

Over, T. M. and Gupta, V.K. (1994) Statistical analysis of mesoscale rainfall: Dependence of random cascade generador on the large-scale forcing. J. Appl. Meteor. 33, 1526-1542.

Over, T. M. and Gupta, V.K. (1996) A space-time theory of mesoscale rainfall using random cascades. J. Geophys. Res. 101, 26319-26331.

Peckham, S. D. (1995) New results for self-similar trees with applications to river networks. Water Resour. Res. 31, 1023-1030.

Peckham, S.D. and Gupta, V.K. (1999) A reformulation of Horton's laws for large river networks in terms of statistical self-similarity. Water Resour. Res. 35, 2763-2777.

Poveda G, Vélez, J.I., Mesa, O.J., Cuartas, A., Barco, J., Mantilla, R.I., Mejía, J.F., Hoyos, C.D., Ramírez, J.M., Ceballos, L.I., Zuluaga, M.D., Arias, P.A., Botero, B.A., Montoya, M.I., Giraldo, J.D., and Quevedo, D.I. (2007) Linking long-term water balances and statistical scaling to estimate river flows along the drainage Network of Colombia. J. Hydrol. Eng. 12, 4-13.

Reggiani, P, Sivapalan, M., Hassanizadeh, S.M. and Gray, W.G. (2001) Coupled equations for mass and momentum balance in a stream network: theoretical derivation and computational experiments. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 457, 157-189.

Richards-Pecou, B. (2002) Scale invariance analysis of channel network width function and possible implications for flood behavior. Hydrol. Sci. J. 47, 387-404.

Rinaldo, A. and Rodriguez-Iturbe, I. (1996) Geomorphological theory of the hydrological response. Hydro. Proc. 10, 803-829.

Rigon, R., Rinaldo, A., and Rodriguez-Iturbe, I., Bras, R.L., and Ijjasz-Vasquez, E. (1993) Optimal channel networks: A framework for the study of river basin morphology. Water Resour. Res. 29, 1635-46.

Robinson, J. S. and Sivapalan, M. (1997) An investigation into the physical causes of scaling and heterogeneity in regional flood frequency. Water Resour. Res. 33, 1045-1060.

Rodriguez-Iturbe, I. and Rinaldo, A. (1997) *Fractal river basins: chance and self-organization*. Cambridge.

Rodriguez-Iturbe, I., Ijjasz-Vasquez, E.J., Bras, R.L., and Tarboton, D.G. (1992) Power law distributions of mass and energy in river basins. Water Resour. Res. 28, 1089-1093.

Rodriguez-Iturbe, I. and Valdes, J.B. (1979) The geomorphologic structure of hydrologic response. Water Resour. Res. 15, 1409-1420.

Schroeder, M. (1991) *Fractals, Chaos, and Power Laws*. W.H. Freeman and Com., New York.

Shreve, R.L. (1966) Statistical law of stream numbers. J. Geol. 74, 17-37.

Shreve, R.L. (1967) Infinite topologically random channel networks. J. Geol. 75, 178-86.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E., O'Connell, P., Oki, T., Pomeroy, J., Schertzer, D., Uhlenbrook, S. and Zehe, E. (2003) IAHS decade on predictions in ungauged basins (PUB), 2003-2012: Shaping an exciting future for the hydrologic sciences. Hydro. Sci. J. 48, 857-880.

Sivapalan, M., Wood, E.F., and Beven, K.J. (1990) On hydrologic similarity, 3, A dimensionless flood frequency model using a generalized geomorphic unit hydrograph and partial area runoff generation. Water Resour. Res. 26, 43-58.

Smith, J. (1992) Representation of basin scale in flood peak distributions. Water Resour. Res. 28, 2993-2999.

Tokunaga, E. (1966) The composition of drainage networks in Toyohira river basin and valuation of Horton's first law (in Japanese with English summary). Geophys. Bull. Hokkaido Univ. 15, 1-19.

Tokunaga, E. (1978) Consideration on the composition of drainage networks and their evolution. Geographical Reports of Tokyo Metropolitan University 13.

Tokunaga, E. (1984) Ordering of divide segments and law of divide segment numbers. Trans. Jpn. Geomorphol. Union, 5, 71-77.

Troutman, B.M. (2005) Scaling of flow distance in random self-similar channel networks. Fractals. 13, 265-282.

Troutman, B.M. and Karlinger, M.R. (1984) On the expected width function for topologically random channel networks. J. Appl. Prob. 21, 836-849.

Troutman, B.M. and Karlinger M.R. (1994) Inference for a generalized Gibbsian distribution on channel networks. Water Resour. Res. 30, 2325-2338.

Troutman, B.M. and Karlinger, M.R. (1998) Spatial channel network models in hydrology. In: O. E. Barndorff-Nielsen, V. K. Gupta, V. Perez-Abreu, and E. C. Waymire (Eds.), *Stochastic Methods in Hydrology: Rainfall, landforms and floods*. Adv. Series Stat. Sci. and Appl. Prob. 7, World Scientific, pp. 85-127.

Troutman, B. M. and Over, T.M. (2001) River flow mass exponents with fractal channel networks and rainfall. Adv. Water Resour. 24, 967-989.

Turcotte, D. L. (1997) *Fractals and Chaos in Geology and Geophysics*, 2nd Edition, Cambridge.

Turcotte, D. L. and Rundle, J.B. (Eds.) (2002) Self-organized complexity in the physical, biological and social sciences, Proc. Natl. Acad. Sci. USA, 99, Suppl. 1, 2463-2465.

Veitzer, S. A., and Gupta, V.K. (2000) Random self-similar river networks and derivations of Horton-type relations exhibiting statistical simple scaling. Water Resour. Res. 36, 1033-1048.

Veitzer, S. A. and Gupta, V.K. (2001) Statistical self-similarity of width function maxima with implications to floods. Adv. Water Resour. 24, 955–965.

Veitzer, S.A., Troutman, B.M. and Gupta, V.K. (2003) Power-law tail probabilities of drainage areas in river basins. Phys. Rev. E. 68, 016123.

Veneziano, D., Moglen, G.E., Furcolo, P. and Iacobellis, V. (2000) Stochastic model of the width function. Water Resour. Res. 36, 1143-1158.

West, G. B. and Brown, J.M. (2004) Life's universal scaling laws. Physics Today 57, 36-42.

9 Investigations of Wave-induced Nonlinear Response of Minor Species with the KBM Averaging Method

Tai-Yin Huang

The Pennsylvania State University, Physics Department, tuh4@psu.edu

Abstract. An analytical solution to the continuity equation of minor species was for a long time obtained only with a linear treatment owing to the fact that secularity would arise when a direct perturbation expansion was applied to the equation. Recently, Huang et al. (2003) applied the Krylov-Bogoliubov-Mitropolsky (KBM) averaging method to remove the higher-order secular terms in the perturbation expansion series, and what remained in the series were the terms that oscillate at frequencies that are an integer multiple of the forcing wave frequency. A vertical drift velocity demonstrated to be the Eulerian drift was required to remove the secular terms. Following the previous work, a coupled chemical-dynamical gravity wave model with the application of the KBM method is presented here to investigate the wave-induced nonlinear response. The significance of the higher-order nonlinear response will be investigated. Results indicate that higher-order terms are significant when the scale height of the minor species is small.

1 Introduction

When a gravity wave propagates obliquely through the atmosphere, several things could happen that would affect the atmosphere. It could transport atmospheric species, thus helping the mixing of chemical reactions and changing the vertical distribution of the constituents (Huang and Hickey 2006a; Huang and Hickey 2006b). It could transfer wave energy and momentum to the mean flow or extract energy from the mean flow, depending on the local stability of the background atmosphere (Fritts 1978; Huang, Hur, Tuan, Li, Dewan, and Picard 1998; Huang, Hickey, Tuan, Dewan, and Picard 2002). The thermal structure and the wind system are thus altered due to this kind of wave-mean flow interactions. When successive wave trains pass through the air, species are perturbed due to the wave forcing. The aforementioned effects are just a few examples of how gravity waves can affect our atmosphere and the environment.

The response of minor species to a gravity wave is, in general, greater than that of the major species because the density profile of minor species like O and O_3 often has a layered structure. Figure 9.1 (from Huang, Hickey, and Tuan 2003) describes a

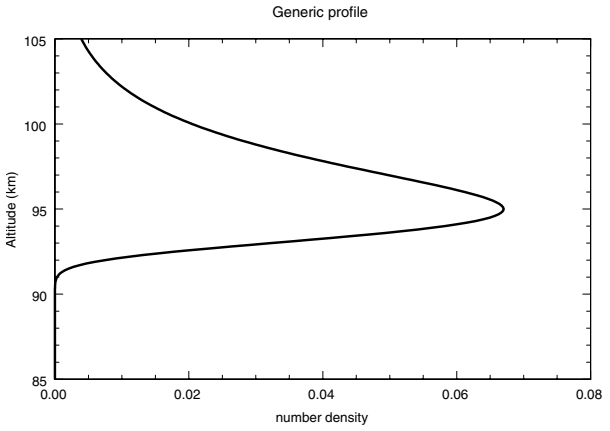


Fig. 9.1. A generic profile peaks at 95 km with a shape factor of 1.5 km. Under appropriate adjustment of its peak value, peak altitude and the shape factor, the profile can simulate a minor species' profile. From Huang et al. (2003)**.**

nominal (generic) vertical profile with a peak altitude at 95 km and a shape factor of 1.5 km as an example. In general, number densities of minor species display such a vertical distribution. The peak altitude, the shape factor and the half-maximum width vary for each different species. Numerous studies (Chiu and Ching 1978; Hatfield, Tuan, Silverman 1981; Gardner and Shelton 1985; Huang et al. 2003) showed that the greater the density gradient of the minor species, the larger the response of the minor species to a small scale gravity wave. It is worth investigating how minor species react to wave forcing because 1) they are directly observable and 2) they are involved in chemical reactions that produce nightglow emissions, which again are observable.

The reason that modeling the response has been largely focused on only the linear effects was because it greatly simplified the problem. The challenge of modeling the nonlinear effects has been that when a direct perturbation expansion technique is used, there will exist a term in the expansion series that is a constant. After time integration, that time constant will result in a term that is linearly proportional to time. The most well known case has to be the nonlinear oscillator where there exists a secularity in the solution if a direct perturbation expansion is performed. Gardner et al. (1985) realized the presence of nonlinearity in the observational data, and they developed an approach to studying the nonlinear effects due to a linear conservative gravity wave by using a presumed solution for the continuity equation. Their solution, however, contains a nonvanishing term when a time average is performed on the quantities (see their equation 44). This nonvanishing term is the secular term, which was not recognized or addressed in their paper. That there remains a nonzero value after the time averaging on the perturbed number density is problematic because it would mean the gravity wave could cause a lasting change like compression

or expansion of the atmosphere. With a nondissipative conservative wave as was used in their paper, such a scenario is not allowed.

In this study, we will first describe the application of Krylov-Bogoliubov-Mitropolsky averaging method in Section 2. The gravity wave model and the wave field (the horizontal and vertical component of a gravity wave) are laid out in Section 3. We then present the coupled chemical-dynamical gravity wave model and the equations for linear and nonlinear response after performing perturbation expansion of the continuity equation in conjunction with the application of the Krylov-Bogoliubov-Mitropolsky method in Section 4. Results focused on the dynamical effects will be presented in Section 5. We conclude in Section 6.

2 The Krylov-Bogoliubov-Mitropolsky Averaging Method

Whenever perturbation expansion is applied to a nonlinear system, secularity exists if higher order terms are retained. Since most of the systems encountered are nonlinear in nature, numerous methods have been developed to treat nonlinear problems (see Jackson (1989)). The Krylov-Bogoliubov-Mitropolsky (KBM hereafter) averaging method is commonly used to remove secularity in a nonlinear system. This method has been used to solve the well-known nonlinear oscillator problem in physics (*Jackson* 1989) where the nonlinear term gives rise to an unphysical linear time divergence in the solution. For a linear oscillator, the frequency is independent of amplitude. However, the frequency of the oscillator becomes a function of amplitude when the oscillatory motion is not linear any more, i.e., when the amplitude becomes large as shown in Fig. 9.2. The application of the KBM method to this nonlinear system is to treat the frequency of the oscillator as a parameter, which will be modified in such a way that secular terms will be removed in the solution.

The parameter to be modified differs for each system as it should be, and it is determined based on the criterion that the unphysical secular terms would be removed after the modification of the parameter. To study the nonlinear response of minor

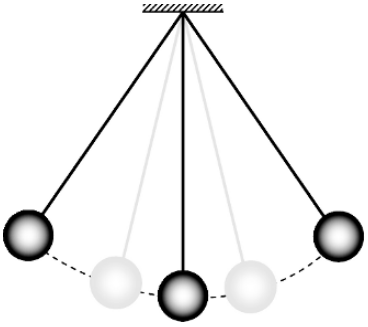


Fig. 9.2. A nonlinear oscillator whose frequency is a function of amplitude.

species to a conservative gravity wave, initially a direct perturbation expansion in the continuity equation was performed to obtain the wave-induced density response, and it was found that the solution of the number density response did contain unphysical secular terms that would increase with time after time integration. Recognizing that secularity did arise when a direct perturbation expansion was applied to the continuity equation, the KBM method was then used in conjunction with the perturbation expansion in order to obtain a solution that will be physical and meaningful. Studies of Walterscheid and Hocking (1991) indicate a small d.c. vertical velocity, nonlinear in nature, that exists in addition to the linear gravity wave vertical velocity. The vertical velocity of gravity wave is thus an obvious choice to be the parameter in this system.

The application of the KBM method is as follows: The vertical velocity of the gravity wave now contains two terms, the linear term and a second-order vertical drift velocity w_d, which is a function of z only. The velocity field of the minor species is assumed to be the same as that of major gases (Chiu et al. 1978), which are tracers of gravity waves. We then apply the perturbation expansion method to the continuity equation. The first-order equation is linear and contains no secular terms, thus the solution can be easily calculated as usual. The second-order equation, however, contains oscillatory terms and constant d.c. secular terms. The secular terms occurring in the higher-order nonlinear equations are removed by setting them equal to the terms containing the vertical drift velocity. The removal of the secular terms from the higher-order continuity equation leaves terms that are purely oscillatory in the equation. The solution thus calculated after time integration is purely oscillatory and describes the response of minor species to a gravity wave with no contamination of secularity.

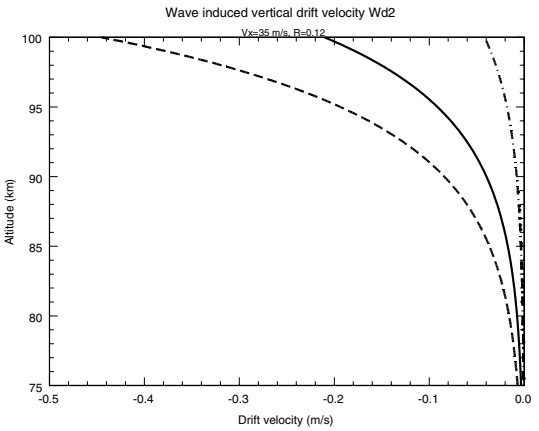


Fig. 9.3. The vertical drift velocity as a function of height. Three waves with the same phase velocity of 35 m/s, but different periods are used. The solid, dot-dashed, and dashed curves are for the wave with a period of 12, 28, and 8 minutes, respectively.

The vertical drift velocity is obtained from the requirement that the secular terms balance the terms containing the vertical drift velocity. The existence of a second-order vertical drift velocity in the gravity wave velocity field not only has its theoretical foundation (Walterscheid et al. 1991) as mentioned above but also receives support from the experimental data (Balsley and Riddle 1984; Coy, Fritts, and Weinstock 1986). This vertical drift velocity is furthered proved in Huang et al. (2003) that it is the Eulerian drift velocity. Figure 9.3 is a plot of the second-order vertical drift velocity as a function of height for three waves with the same horizontal phase velocity (35 m/s) but different wave periods. The figure reveals that the second-order vertical drift velocity has a magnitude of tens of centimeters per second, and its direction is downward since it has negative values. The downward direction and the magnitude of the drift velocity further support the claim made by Coy et al. (1986) that the mean downward vertical velocity observed by Balsley et al. (1984) could be explained by the downward Eulerian mean velocity.

3 The Gravity Wave Model

Here we describe the gravity wave model adapted from the Hines' linear gravity wave model (1960) as part of the chemical-dynamical gravity wave model for the study. Assuming the motion to be two dimensional after proper coordinate transformation, the linearized hydrodynamic equations in an inviscid flat windy atmosphere are:

$$i\Omega\rho' + w'\frac{d\rho_0}{dz} - ik_x\rho_0 u' + \rho_0\frac{\partial w'}{\partial z} = 0 \tag{1}$$

$$i\Omega\rho_0 u' + \rho_0\frac{dV_h}{dz}w' = ik_x p' \tag{2}$$

$$i\Omega\rho_0 w' = -g\rho' - \frac{\partial p'}{\partial z} \tag{3}$$

$$i\Omega p' + w'\frac{dp_0}{dz} = i\Omega c^2\rho' + c^2 w'\frac{d\rho_0}{dz}, \tag{4}$$

In the equations, ρ', p', u', w' are density, pressure, horizontal and vertical velocity perturbations of a gravity wave, respectively. The background atmospheric density ρ_0 and the background atmospheric pressure p_0 are assumed to be a function of z only. The horizontal background wind profile $V_h(z)$ is assumed to vary slowly in the vertical direction. $\Omega(z) = \omega - k_x V_h(z)$ is the Doppler–shifted frequency, ω is the intrinsic wave frequency, k_x is the horizontal wave number, c is the speed of sound and g is the gravitational acceleration. As it shows, the Doppler-shifted frequency is a function of height because the horizontal background wind is assumed to vary in height.

The above four coupled equations can be reduced to two coupled equations (Wang and Tuan 1988; Makhlouf, Picard, Winick, and Tuan 1998) that contain only two variables:

$$\frac{\partial \Delta\overline{w}}{\partial z}+\left(\eta+\frac{k_x}{\Omega}\frac{dV_h}{dz}\right)\Delta\overline{w}=i\left(\frac{k_x^2}{\Omega^2}-\frac{1}{c^2}\right)\Omega\Delta\overline{p} \tag{5}$$

$$\frac{\partial \Delta\overline{p}}{\partial z}-\eta\Delta\overline{p}=i\left(\frac{\omega_b{}^2-\Omega^2}{\Omega}\right)\Delta\overline{w} \tag{6}$$

where $\Delta\overline{w}=\sqrt{\rho_0}w'$, $\Delta\overline{p}=p'/\sqrt{\rho_0}$, $\eta=-\left(1-\frac{\gamma}{2}\right)\frac{g}{c^2}$ is the negative of Eckart's coefficient (Gossard and Hooke 1975), γ is the ratio of specific heats and $\omega_b{}^2=\frac{(\gamma-1)g^2}{c^2}$ is the square of Brunt frequency. These two coupled first–order differential equations can be combined, after elimination of one of the variables, to become a single second–order differential equation that describes the sinusoidal wave behavior in the vertical direction. That is

$$\frac{\partial^2 \Delta\overline{w}}{\partial z^2}+\left[\left(\frac{k_x{}^2}{\Omega^2}-\frac{1}{c^2}\right)\left(\omega_b{}^2-\Omega^2\right)-\eta^2\right]\Delta\overline{w}=0 \tag{7}$$

The terms in the brackets can be readily identified as the square of the vertical wave number k_z. Plane wave solution in the (*x*,*t*) domain is assumed for all the perturbed quantities, and the horizontal and vertical components of gravity wave velocity can be derived by using Eq. 2 and Eq. 6. They are

$$u'(x,z,t)=\left|u(z)\right|e^{i(\omega t-k_x x)}\left[e^{i(k_z z+\pi-\phi)}+R\,e^{-i(k_z z-\phi)}\right] \tag{8}$$

$$w'(x,z,t)=\left|w(z)\right|e^{i(\omega t-k_x x)}\left[e^{ik_z z}+R\,e^{-ik_z z}\right] \tag{9}$$

where

$$\left|u(z)\right|=A\omega k_x c^2\sqrt{k_z{}^2+\eta^2}\,e^{z/2H}(\omega^2-c^2k_x{}^2)/(\Omega^2-c^2k_x{}^2) \tag{10}$$

$$\left|w(z)\right|=A\omega\left(\omega^2-c^2k_x{}^2\right)e^{z/2H} \tag{11}$$

$$k_z=\sqrt{\left(\frac{k_x{}^2}{\Omega^2}-\frac{1}{c^2}\right)\left(\omega_b{}^2-\Omega^2\right)-\eta^2} \tag{12}$$

$$\phi=\tan^{-1}\left(\eta/k_z\right) \tag{13}$$

Where A is the wave amplitude and $H=c^2/\gamma g$ is the atmospheric scale height. Because the wind profile is not constant in height, waves can be reflected. So we retain the reflection coefficient R in our wave solution to take such a possibility into account.

4 The Coupled Chemical-Dynamical Gravity Wave Model

The continuity equation that governs the evolution of minor species *i* is

$$\frac{\partial n_i}{\partial t} = P_i - n_i L_i - \nabla \cdot \left(n_i \vec{V}_i \right) = \Delta \left(PL \right)_i - \nabla \cdot \left(n_i \vec{V}_i \right) \tag{14}$$

where n is the total number density that comprises the background number density $n_0(z)$ and the number–density perturbation $n'(x,z,t)$, i.e., $n(x,z,t) = n_0(z) + n'(x,z,t)$, P and L are the volumetric minor species production rate and the minor species' loss frequency, respectively. Here the subscript i in the equation denotes the minor species i employed. We use $\Delta(PL)$ as a shorthand for the net difference between the production and loss of the minor species. The velocity field $\vec{V}$ of the minor species, in the presence of a horizontal wind, is

$$\vec{V} = [u(x,z,t) + V_h(z)]\hat{i} + [w(x,z,t) + w_d(z)]\hat{k} \tag{15}$$

where $\hat{i}$ and $\hat{k}$ are unit vectors in the x– and z– direction, respectively. The wave–induced drift velocity w_d is a function of z only. In Chiu et al. (1978) and Hatfield et al. (1981), the velocity field of the minor species is assumed to be the same as the linear gravity–wave velocity field. This is acceptable when only the linear response is considered. However, when dealing with the nonlinear (higher–order) response as in this study, secularity which is a nonlinear effect would arise. The appearance of the secularity is not physical owing to the fact that the wave, be it conservative or dissipative, should not induce long–term effects like expansion or contraction in the atmosphere (see Walterscheid et al. (1991) and Huang et al. (2003) for the extensive discussion and derivation based on this argument). It is well known that secular terms (which will produce solutions linearly proportional to time) will arise in the nonlinear system when a direct perturbation expansion is employed. To remove the secularity, we use the KBM averaging method (see Huang et al. (2003) and Jackson (1989)). The advantage of the KBM averaging method is that by modifying some physical parameters in the system, the secular terms will vanish and well–behaved solutions from the perturbation expansion will result. A wave–induced drift velocity existing in the vertical direction is thus chosen to be the parameter, and it was found by Huang et al. (2003) that this drift velocity is the Eulerian drift velocity.

Once we are assured that the secular terms will be removed by employing the KBM method, we can then apply the perturbation expansion to the continuity equation. We let $\Psi_i = \dfrac{n_i - n_{0,i}}{n_{0,i}}$, the ratio of the density perturbation of minor species i to

its background unperturbed number density, and insert the velocity field in the continuity equation (Eq. 14) to obtain

$$\frac{\partial \Psi_i}{\partial t} = (n_{0,i})^{-1}\Delta(PL)_i - (u+V_h)\frac{\partial \Psi_i}{\partial x} - (w+w_{d,i})\frac{\partial \Psi_i}{\partial z} - S_i(1+\Psi_i) - (1+\Psi_i)(n_{0,i})^{-1}\frac{d}{dz}(n_{0,i}w_{d,i}) \quad (16)$$

where

$$S_i = \frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} + (n_{0,i})^{-1}\frac{dn_{0,i}}{dz}w \quad (17)$$

The response of minor species i in the perturbation expansion series in powers of λ (where λ is the expansion coefficient that keeps track of the orders) can be expressed as

$$\Psi_i = \lambda\psi_{1,i} + \lambda^2\psi_{2,i} + \lambda^3\psi_{3,i} + \cdots \quad (18)$$

It should be noted that u and w are treated as first-order terms because a small-scale gravity wave is assumed in this study. The vertical drift velocity can be expanded in a series in a similar fashion. However, the drift velocity w_d exists only in the even–order terms (Huang et al. 2003). All odd–order terms are automatically omitted because they do not contribute to fluctuations in the number densities of minor species. The series becomes:

$$w_{d,i} = \lambda^2 w_{d2,i} + \lambda^4 w_{d4,i} + \cdots \quad (19)$$

Substituting Eqs. 18 and 19 into Eq.16 and collecting terms by the same order in powers of λ, we get equations up to any order. The first-order and the second-order equations are

$$\frac{\partial \psi_{1,i}}{\partial t} = -V_h\frac{\partial \psi_{1,i}}{\partial x} - S_i + (n_{0,i})^{-1}\Delta(PL)_{1,i} \quad (20)$$

and

$$\frac{\partial \psi_{2,i}}{\partial t} = -V_h\frac{\partial \psi_{2,i}}{\partial x} - u\frac{\partial \psi_{1,i}}{\partial x} - w\frac{\partial \psi_{1,i}}{\partial z} - S_i\psi_{1,i} - (n_{0,i})^{-1}\frac{d}{dz}(n_{0,i}w_{d2,i}) + (n_{0,i})^{-1}\Delta(PL)_{2,i} \quad (21)$$

As is shown in the above two equations, the vertical drift velocity does not exist in the first-order equation. However, all higher order equations do contain the vertical drift velocity in the equations. In general, the mth-order equation can be expressed as the following

$$\frac{\partial \psi_{m,i}}{\partial t} = -V_h \frac{\partial \psi_{m,i}}{\partial x} - u\frac{\partial \psi_{m-1,i}}{\partial x} - w\frac{\partial \psi_{m-1,i}}{\partial z} - S_i \psi_{m-1,i}$$
$$-\sum_{k=2}^{m} w_{dk,i} \frac{\partial \psi_{m-k}}{\partial z} - \sum_{k=2}^{m} \psi_{m-k} (n_{0,i})^{-1} \frac{d}{dz}\left(n_{0,i} w_{dk,i}\right) + (n_{0,i})^{-1} \Delta\left(PL\right)_{m,i} \quad (22)$$

After expansion and collecting terms, the even-order equation can be cast into the following form

$$\mathrm{Re}(\frac{\partial \psi_{n,i}}{\partial t}) = oscillatory\ terms + (\mathrm{sec}ular\ terms + w_{dn,i}\ terms) \quad (23)$$

From Eq.23, it is obvious that the terms in the parentheses need to cancel out in order to have the solutions that are purely oscillatory. That is,

$$\mathrm{Re}(\frac{\partial \psi_{n,i}}{\partial t}) = oscillatory\ terms \quad (24a)$$

and

$$\mathrm{sec}ular\ terms + w_{dn,i}\ terms = 0 \quad (24b)$$

The response of the minor species can then be obtained by a direct time integration of Eq.24a. The vertical drift velocity can be calculated by the requirement that secular terms balance the terms containing the vertical drift velocity. It is found that the first two-order responses make the most significant contribution to the fluctuations in the number densities of the minor species.

5 Results

A Chapman-like function is used to simulate the unperturbed number density profile of minor species in the current study. The expression of a Chapman-like function for the unperturbed number density of minor species (j) is $n_{0j}(z) = A_j e^{F_j}$ where $F_j(z) = 1 - \frac{z - z_j}{h_j} - e^{-\left(\frac{z-z_j}{h_j}\right)}$ and where A_j is the peak value of the background number density, z_j is the peak altitude, h_j is the shape factor, which is the scale height of the minor species. We used a generic profile and several species (H, O_3 and OH) for the investigation. The reason for using a generic profile is that we can adjust the parameter like the shape factor and see how the hypothetical species responds to a wave with a different scale height. For the current study, results focusing on the dynamics (equivalent to ignoring chemistry) will be presented. A small-scale, short-period gravity wave is used to investigate the dynamical and transport effects on the minor species.

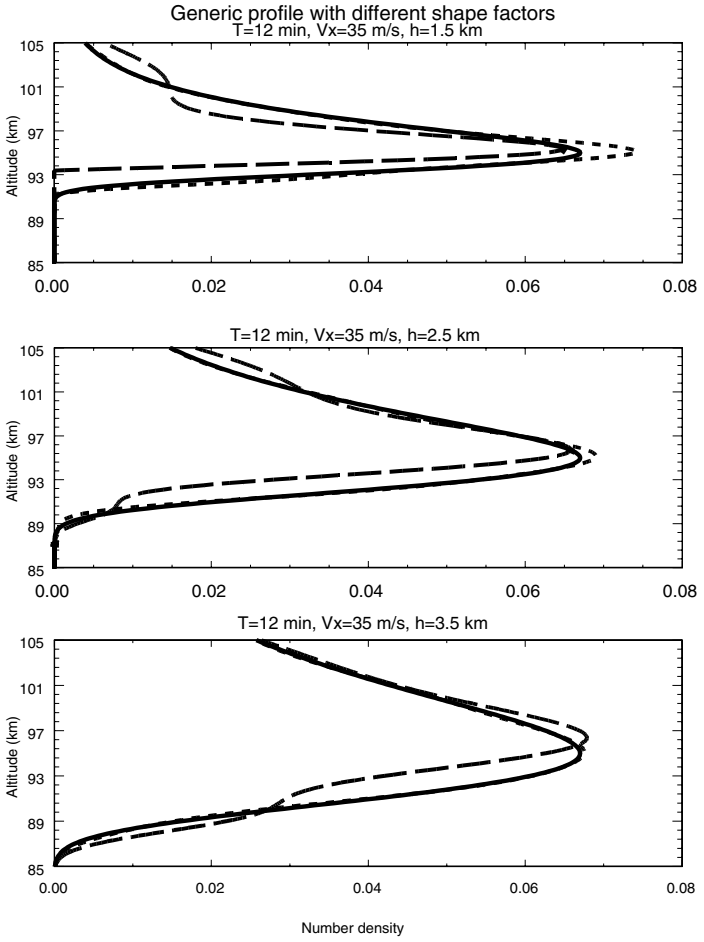


Fig. 9.4. The spatial vertical profile of the response of a generic profile with different shape factors: 1.5 km, 2.5 km and 3.5 km in the top, middle and bottom panels, respectively, at a time when t=0. The solid curve is the background density, dashed curve is background density plus the first-order response, and the dotted curve is the background density plus the second-order response. From Huang et al. (2003).

Figure 9.4 shows the vertical profiles of the response of an arbitrary species described by a generic profile with three different minor species' scale heights (shape factor): 1.5 km in the top panel, 2.5 km in the middle panel and 3.5 km in the bottom panel. The gravity wave used in this figure has a wave period of 12 minutes and a horizontal phase velocity of 35 m/s and is assumed to be a freely traveling wave, i.e.,

the reflection coefficient is set to zero. The envelope of the wave amplitude is normalized to produce a 5% variation in the density of the major species at 90 km altitude. The solid curve denotes the unperturbed background number density and the dashed curve is the background number density plus the first-order response. The dotted curve denotes the background number density plus the second-order response. Comparing these three plots, it is apparent that the smaller the minor species' scale height the greater the second-order response and the broader the profile the less important the higher-order terms. We calculated the spatial response of H, O_3 and OH and found that the second-order spatial response for these three species are smaller and thus less important than those shown in Fig. 9.4, mainly because the scale heights of these three species are considerably larger than any of the values used for the generic profile.

Figure 9.5 shows the temporal response of a generic profile with a fixed minor species' scale height (a 2.5 km value is used) at three different altitudes. We plotted

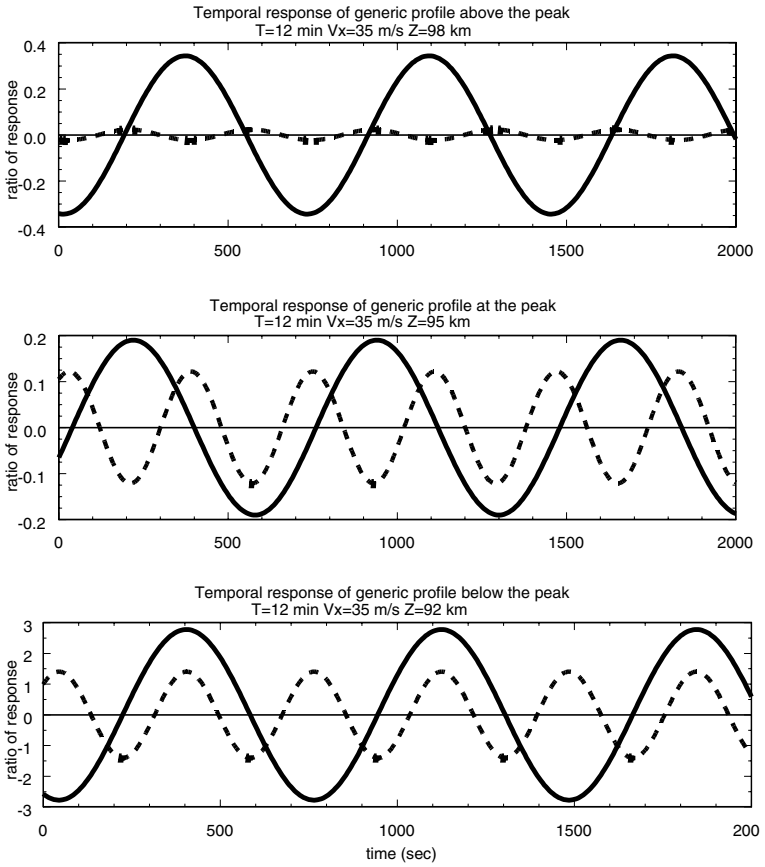


Fig. 9.5. The temporal response of a generic profile with a shape factor of 2.5 km is plotted at three altitudes: above the peak, at the peak and below the peak in the top, middle and bottom panels, respectively. From Huang et al. (2003).

the response above the peak in the top panel, at the peak in the middle panel and below the peak in the bottom panel. The solid curves are for the ratio of first-order response to the background number density while the dashed curves are for the ratio of second-order response to the background number density. Inspection on these plots reveals that the first- and second-order response of an arbitrary minor species are much greater below the peak than those at or above the peak. Because the second-order response is large below the peak, the response is clearly nonlinear. In addition, the second-order response is in phase with the first-order response every other cycle below the peak. The superposition of the first- and second-order terms results in an even more pronounced response. Thus higher-order terms are essential for correctly describing the response of minor species to wave forcing below the peak. The second-order response above the peak, however, is diminutive that the negligence of higher-order terms can be tolerated.

A wide range of wave parameters are used in the calculation of the response to see how sensitive the minor species responds to different wave forcing. Figure 9.6 is a plot of the first and second order wave-induced response of a generic profile as a function of period at 87-km altitude. The shape factor of the generic profile is 2.5 km and the horizontal wavelength of the wave is set at 50 km. The actual magnitude of the second order response is ten times smaller than what is shown in the figure. The figure shows that the magnitude of the first and second order response is not heavily dependent of the wave periods except very near the Brunt period. Waves with small wave periods (longer vertical wavelengths) seem to produce a much larger response.

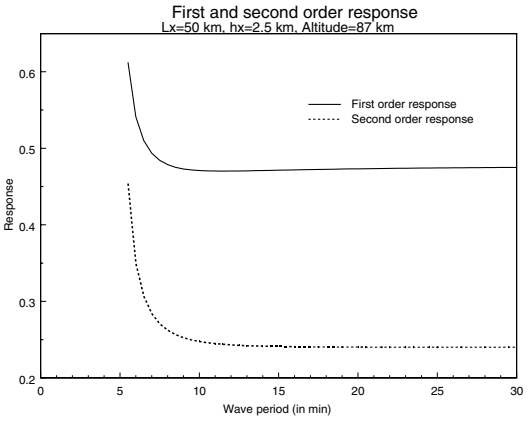


Fig. 9.6. First and second order response of a generic profile induced by waves with a horizontal wavelength 50 km at an altitude of 87 km. The solid curve is for the first-order response and the dashed curve for the second-order response. Note that the actual magnitude of the second-order response is ten times smaller than is shown in the figure. From Huang et al. (2003).

6 Conclusion

Our results clearly indicate the importance of the nonlinear response below the peak altitude of minor species. When the scale height of minor species is small, nonlinearity is expected to play a significant role in the response of the species to a small-scale gravity wave. As a result, a linear analysis on the observational data may not be sufficiently adequate to extract accurate wave information from the data. Therefore, care must be taken when analyzing data that might contain nonlinearity.

7 Acknowledgments

This work was supported by NSF Grant ATM-0406585 to The Pennsylvania State University. Figures 9.1, 9.4, 9.5 and 9.6 are reproduced by permission of American Geophysical Union from an article by Tai-Yin Huang, Michael Hickey and Tai-Fu Tuan, On nonlinear response of minor species with a layered structure to gravity waves, Journal of Geophysical Research, Vol. 108, No. A5, 1173, doi:10.1029/2002JA009497, 2003. Copyright 2003 American Geophysical Union.

References

Chiu, Y. T. and Ching, B. K. (1978) The response of atmospheric and lower ionospheric layer structures to gravity waves, *Geophys. Res. Lett*., 5, 539-542.

Coy, L., Fritts, D. C. and Weinstock, J. (1986) The Stokes Drift due to Vertically Propagating Internal Gravity Waves in a Compressible Atmosphere, *J. Atmos. Sci., 43*, 2636-2643.

Fritts, D. C. (1978) The nonlinear gravity-wave critical-layer interaction, *J. Atmos. Sci.*, 35, 397-413.

Gardner, C. S. and Shelton, J. D. (1985) Density response of neutral atmospheric layers to gravity wave perturbations, *J. Geophys. Res.,* 90, 1745.

Gossard, E. E. and Hooke, W. H. (1975) Waves in the Atmosphere, Elservier, New York.

Hatfield, R., Tuan, T. F. and Silverman, S. M. (1981) On the effects of atmospheric gravity waves on profiles of *H*, *O3* and *OH* emission, *J. Geophys. Res*., 86, 2429-2437.

Hines, C. O. (1960) Internal atmospheric gravity waves at ionospheric heights, *Can. J. Phys*., 38, 1441-1481.

Huang, T.-Y. and Hickey, M. P. (2006a) On the latitudinal variations of the non-periodic response of minor species induced by a dissipative gravity-wave packet in the MLT region, accepted for publication in *J. Atmos. Solar-Terr. Phys.*.

Huang, T.-Y. and Hickey, M. P. (2006b) Secular variations of OH nightglow emission and of the OH intensity-weighted temperature induced by gravity-wave forcing in the MLT region, submitted to *Advances in Space Research*.

Huang, T-Y., Hickey, M. P. and Tuan, T. F. (2003) On nonlinear response of minor species with a layered structure to gravity waves, *J. Geophys. Res*., 108, doi:10.1029/ 2002JA009497.

Huang, T.-Y., Hickey, M. P., Tuan, T. F., Dewan, E. M. and Picard, R. H. (2002) Further investigations of a mesospheric inversion layer observed in the ALOHA-93 Campaign, *J. Geophys. Res.*, V107, 4408, doi:10.1029/2001JD001186.

Huang, T.-Y., Hur, H., Tuan, T. F., Li, X., Dewan, E. M. and Picard, R. H. (1998) Sudden narrow temperature inversion-layer formation in ALOHA-93 as a critical-layer-interaction phenomenon, *J. Geophys. Res.*, V103, 6323-6332.

Jackson, E. A.(1989) Perspectives of nonlinear dynamics, Cambridge University Press.

Makhlouf, U. B., Picard, R. H., Winick, J. R. and Tuan, T. F. (1998) A model for the response of the atomic oxygen 557.7 nm and the OH Meinel airglow to atmospheric gravity waves in a realistic atmosphere, *J. Geophys. Res*., 103, 6261.

Walterscheid, R. L. and Hocking, W. K. (1991) Stokes Diffusion by Atmospheric Internal Gravity Waves, *J. Atmos. Sci.* 48, 2213-2230.

Wang, D. Y. and Tuan, T. F. (1988) Brunt-Doppler ducting of small-period gravity waves, *J. Geophys. Res*., 93, 9916-9926.

10 ENSO Signal Propagation Detected by Wavelet Coherence and Mean Phase Coherence Methods

S. Jevrejeva[1], J.C. Moore[2], and A. Grinsted[2,3]

[1] Proudman Oceanographic Laboratory, Liverpool, UK, sveta@pol.ac.uk
[2] Arctic Centre, University of Lapland, Rovaniemi, Finland, jmoore@ulapland.fi
[3] Dept. of Geophysics, University of Oulu, Oulu, Finland, ag@glaciology.net

Abstract. We present observational evidence of the dynamic linkages between ENSO and Northern Hemisphere (NH) ice conditions over the past 135 years. Using Wavelet Transform (WT) we separate statistically significant components from time series and demonstrate significant co-variance and consistent phase differences between NH ice conditions and the Arctic Oscillation and Southern Oscillation indices (AO and SOI) at 2.2, 3.5, 5.7 and 13.9 year periods. To study the phase dynamics of weakly interacting oscillating systems we apply average mutual information and mean phase coherence methods. Phase relationships for the different frequency signals suggest that there are several mechanisms for distribution of the 2.2-5.7 year and the 13.9 year signals. The 2.2- 5.7 year signals, generated about three months earlier in the tropical Pacific Ocean, are transmitted via the stratosphere, and the Arctic Oscillation (AO) mediating propagation of the signals. In contrast the 13.9 year signal propagates from the western Pacific as eastward propagating equatorial coupled ocean waves, and then fast boundary waves along the western margins of the Americas to reach both polar regions, and has a phase difference of about 1.8-2.1 years by the time it reaches the Arctic.

1 Motivation

The impact of El Niño Southern Oscillation (ENSO) on global climate is ubiquitous, irregular, and largely unpredictable as both oceanic and atmospheric pathways are involved, resulting in considerable weakening of signal/noise ratio. There considerations motivated us to apply new nonlinear techniques (average mutual information and mean phase coherence methods) in order to detect complex and sufficiently weak nonlinear interactions. Several studies indicate that the impact of ENSO on NH climate is not robust, but is more plausibly seen during winter (Pozo-Vázquez, Esteban-Parra, Rodrigo, and Castro-Diez, 2001). Most evidence on teleconnections comes from model simulations (e.g. Merkel and Latif, 2002) or from 50 years of reanalysis data (Ribera and Mann, 2002). We use observational data of NH ice conditions to trace the ENSO signal propagation; since ice extent is an integrated

parameter of winter seasons and in addition acts as a nonlinear filter for 2-13 year oscillations (Jevrejeva and Moore, 2001). Previously oscillations with 2.2, 3.5, 5.7 and 13.9 year periodicities have been detected in ice conditions time series and associated with similar signals in SOI/Niño3 and the NH annular mode (NAM) (Jevrejeva and Moore, 2001; Jevrejeva, Moore and Grinsted, 2003). Results are consistent with quasi-biennial (QB) and quasi- quadrennial (QQ) signals detected by (Gloersen, 1995; Venegas and Mysak, 2000) in shorter time series of ice conditions in polar regions and with signals in SST and sea level pressure (SLP) anomalies over the Pacific Ocean (Huang, Higuchi and Shabbar, 1998; Torrence and Webster, 1999; White and Tourre, 2003, Ribera and Mann, 2002).

The main aim of this paper is to investigate the dynamic connections between the ENSO and NH ice conditions; and to estimate how strong future evolution of one system's phase depends on the other system's phase.

2 Methods

We used the Continues Wavelet Transform (CWT) (Foufoula-Georgiou and Kumar, 1995) to determine both the dominant modes of variability and how those modes vary in time. Two useful wavelets are the Morlet (Foufoula-Georgiou and Kumar, 1995), defined as

$$\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\frac{1}{2}\eta^2}, \tag{1}$$

and the Paul (Torrence and Compo, 1998):

$$\psi_0(\eta) = \frac{2^m i^m !}{\sqrt{\pi(2m)!}} (1 - i\eta)^{-(m+1)}, \tag{2}$$

where $\omega_0$is dimensionless frequency and η is dimensionless time, and m is the order, taken as 4 here. The idea behind the CWT is to apply the wavelet as a band pass filter to the time series. The wavelet is stretched in time, t, by varying its scale (s), so that $\eta = s \cdot t$, and normalizing it to have unit energy. The Morlet wavelet (with ω_0=6) provides a good balance between time and frequency localization and is a good choice for feature extraction. For broad band pass filtering applications, we use the Paul as this is much less localized in frequency space. The CWT of a time series X, $\{x_n, n=1,\ldots,N\}$ with uniform time steps δt, is defined as the convolution of x_n with the scaled and normalized wavelet.

$$W_n^X(s) = \sqrt{\frac{\delta t}{s}} \sum_{n'=1}^{N} x_{n'} \psi_0 \left[(n' - n) \frac{\delta t}{s} \right]. \tag{3}$$

The complex argument of $W^X{}_n(s)$ can be interpreted as the phases of $X\{\phi_1 \ldots, \phi_n\}$.

We define (Grinsted, Moore and Jevrejeva, 2004) the wavelet coherence of two time series X and $Y\{y_1, \ldots y_n\}$ as

$$R_n^2(s) = \frac{\left|S\left(s^{-1}W_n^{XY}(s)\right)\right|^2}{S\left(s^{-1}\left|W_n^X(s)\right|^2\right)\cdot S\left(s^{-1}\left|W_n^Y(s)\right|^2\right)}, \tag{4}$$

where S is a smoothing operator. Notice that this definition closely resembles that of a traditional correlation coefficient, and it is useful to think of the wavelet coherence as a localized correlation coefficient in time frequency space. The smoothing operator $S(W) = S_{scale}(S_{time}(W_n(s)))$, where S_{scale} denotes smoothing along the wavelet scale axis and S_{time} smoothing in time. We apply Monte Carlo methods using a red noise model based on the autocorrelation functions of the two time series to provide statistical significance (Grinsted et al., 2004).

We also consider that nonlinear interactions between the two time series may be chaotic. Causality relationships are analysed using average mutual information (Papoulis, 1984) and the mean phase coherence (Mokhov and Smirnov, 2006) methods. Both these methods rely on the phase expression of the time series derived from (3) with the Paul wavelet (2) of the desired Fourier wavelength, $\lambda = 4\pi s/(2m+1)$, (Torrence and Compo, 1998). The broadband Paul wavelet allows signals that are relatively aperiodic to be included in the analysis. We utilize the average mutual information, *I(X,Y)*, between the two series. In our case we are interested in causative relations, so it is appropriate to measure the *I(X,Y)*, between their respective phases ϕ , and θ.

$$I(X,Y) = \frac{1}{\log_2 B}\sum_{y\in Y}\sum_{x\in X} p(\phi,\theta)\log_2\frac{p(\phi,\theta)}{f(\phi)g(\theta)} \tag{5}$$

where p is the joint probability distribution function of X and Y, and f and g are the marginal probability distribution functions of X and Y respectively, I is normalized by B, the number of histogram bins used to construct f and g.

Another measure of coherence between the two time series is the angle strength of the phase angle difference between the series, also known as the mean phase coherence, ρ:

$$\rho = \frac{1}{N}\sqrt{\left[\sum_{t=1}^{N}\cos(\phi_t - \theta_t)\right]^2 + \left[\sum_{t=1}^{N}\sin(\phi_t - \theta_t)\right]^2} \tag{6}$$

We can search for the optimum relative phase delay between the two series by lagging one time series relative to the other by a phase lag, Δ, in both (5) and (6).

3 Data

We used a 135-year time series of the monthly SOI (Ropelewski and Jones, 1987), as the atmospheric component of ENSO and the Niño3 SST index (1857-2001) (Kaplan, Cane, Kushnir, Clement, Blumenthal and Rajagopalan, 1998), defined as the monthly SST averaged over the eastern half of the tropical Pacific (5ºS- 5ºN, 90º - 150ºW) as the oceanic part. To ensure the robustness of the results, we repeated the analyses for the monthly Niño1+2 (0º - 10ºS, 90º -80ºW) and monthly Niño3.4 (5ºS-5ºN, 170º -120ºW) time series (Kaplan et al., 1998). Northern Hemisphere annular mode (NAM) was represented by the monthly AO index based on the pressure pattern (1899-2001) (Thompson and Wallace, 1998) and the extended AO index (1857-1997) based on the pattern of surface air temperature anomalies (Thompson and Wallace, 1998).

Ice condition in NH were represented by maximum annual ice extent in the Baltic Sea (BMI), 1857-2000 (Seinä and Palosuo, 1996), April ice extent in the Barents Sea (BE: Eastern part 10º- 70ºE; and BW: Western part 30ºW- 10ºE), 1864-1998 (Venje, 2001) and the date of ice break-up at Riga since 1857 (Jevrejeva and Moore, 2001). Global SST since 1854 were taken from the ERSST (Smith and Reynolds, 2003) 2°x2° dataset down-sampled to yearly values.

4 Results

Results from wavelet coherence (Figure 10.1) show that relationship between ENSO and NH ice conditions is relatively weak and varies considerably with time. Figures 10.1 a,b demonstrate that the relationship between ice conditions and AO is not stationary and the influence of AO has generally increased over the last 60 years over a broadening spectrum of periods. There is evidence of a shift in the period of maximum coherence from the 2.2- 5.2 to the 3.5-7.8 and 12-20 year band for the Baltic sea after 1920. Slightly weaker ENSO influence on the variability of ice conditions associated with 2.2- 3.5, 5.2-7.8 and 12-20 year signals, with noticeable increase in the 12-20 year band after 1920 is shown on Figs. 10.1 c,d.

Robustness of coupling via 2.2-5.7 year signals is examined by phase dynamics modeling. For this analysis we use the average mutual information (5) and mean phase coherence (6). We firstly filter both time series with a Paul wavelet centred on 3 and 5 years (Figure 10.2). The peak in I and ρ is at 0-3 months with SOI leading Niño3 and Barents sea ice conditions, Niño3 leading Barents sea ice conditions for the 3-year signals (Jevrejeva et al, 2004). The ENSO influence via 12-20 year signals is maximal for 2-5 year lag (not shown here).

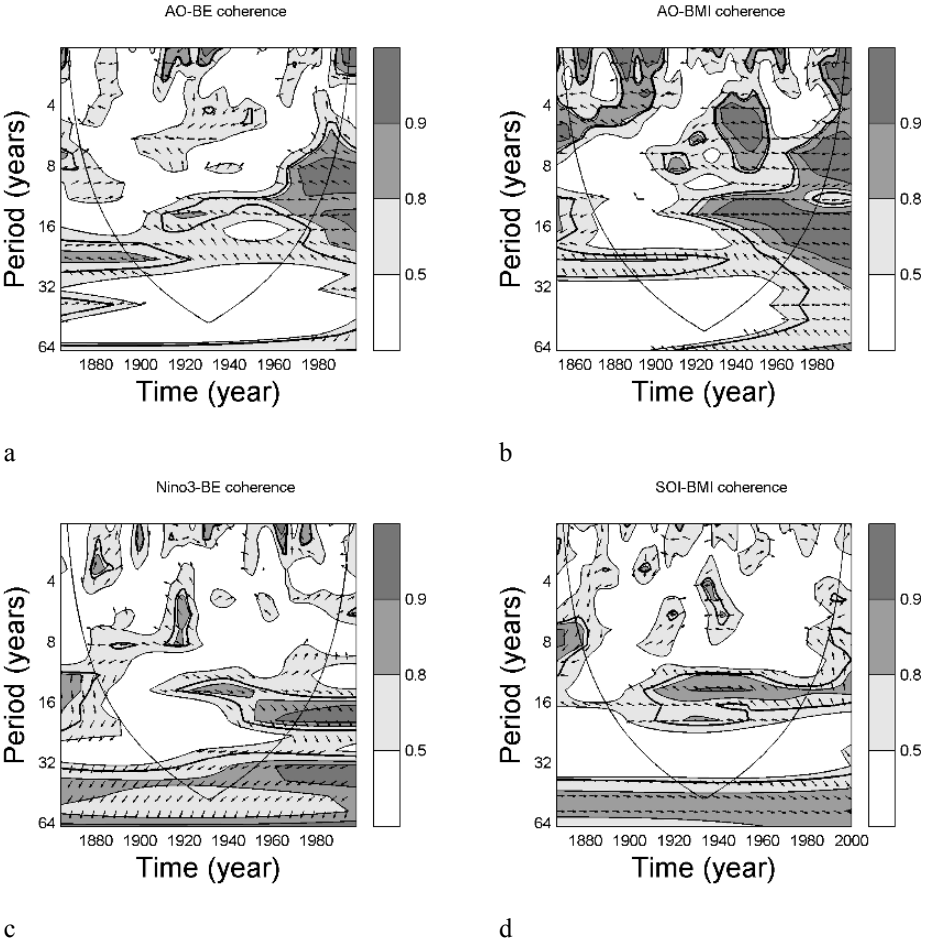


Fig. 10.1. The wavelet coherency and phase difference between the AO/Barents sea ice extent (a). Contours are wavelet squared coherencies. The vectors indicate the phase difference (a horizontal arrow pointing from left to right signifies in-phase and an arrow pointing vertically upward means the second series lags the first by 90 degrees (i.e. the phase angle is 270º); b) the same for the AO/ice extent in the Baltic Sea; c) the same for the Niño3/ice extent in the Barents Sea; d) the same for the SOI/ice extent in the Baltic Sea.

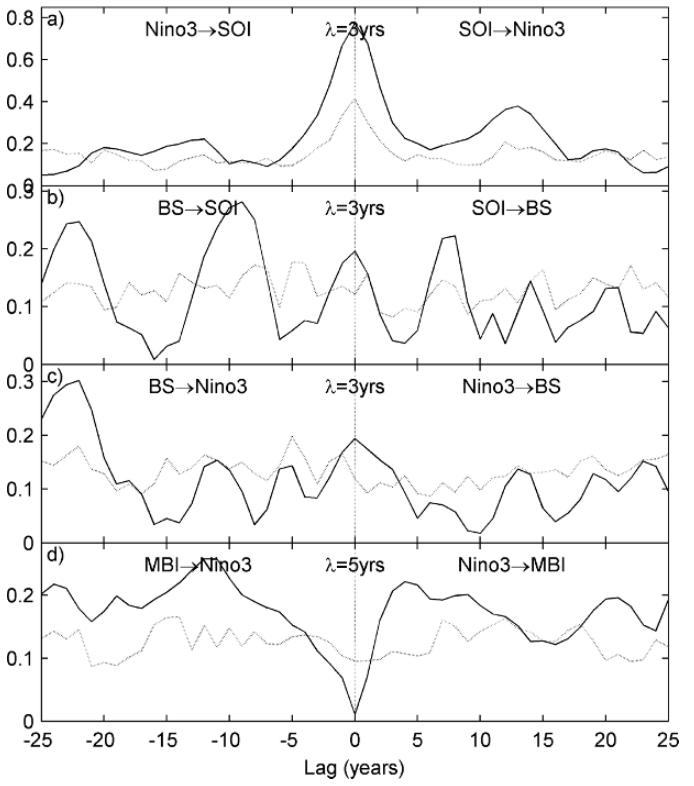
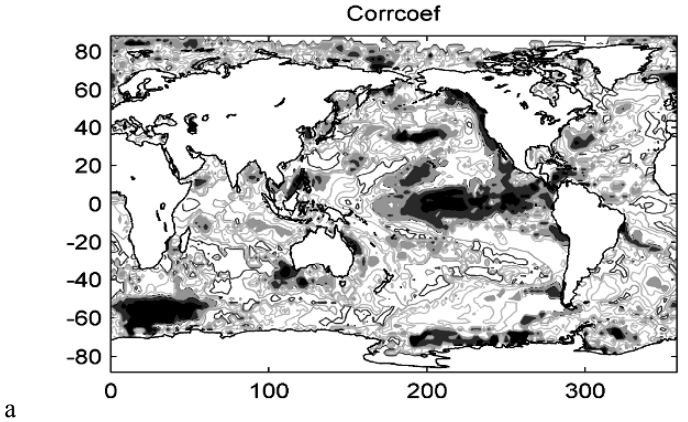


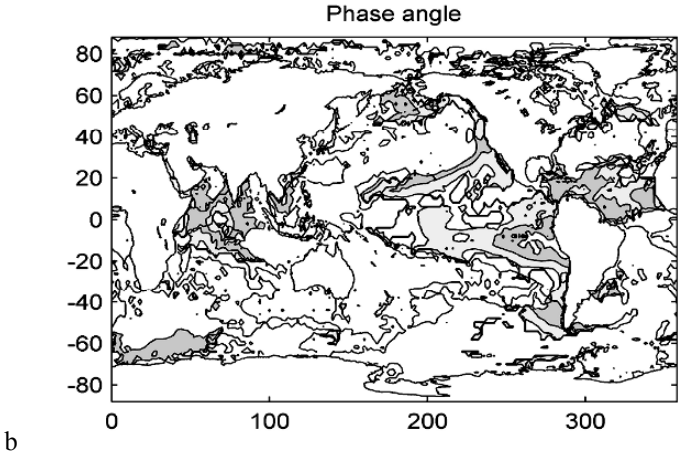
Fig. 10.2. Relationships between time series that have been Paul wavelet filtered with centre frequency λ, expressed as average mutual information, (I), (dotted curves), and mean phase coherence, ρ, (solid line) as a function of the phase lag (Δ) between the series for a) SOI and Niño3, b) SOI and Barents sea ice conditions, c) Niño3 and Barents sea ice conditions, d) Niño3 and Baltic sea ice conditions.

The different phase relationship between signals with periodicities of 2.2-5.7 and those of 12-20 years suggests different mechanisms of signal propagation from the equatorial Pacific Ocean to the polar regions for the decadal-scale signals and for the sub-decadal ones. Similar results are revealed by Jevrejeva et al. (2004). The 2.2- 5.7 year signals are most likely transmitted via the stratosphere, and the AO mediating propagation of the signals, through coupled stratospheric and tropospheric circulation variability that accounts for vertical planetary wave propagation. The connection between the QB oscillations (2.2-3.5 years) detected in ice conditions time series and in tropical forcing, has been described by Baldwin and Dunkerton (2001), as extra-tropical wave propagation, affecting breakdown of the wintertime stratospheric polar

vortices. Since the AO may be interpreted as a physical phenomenon associated with the structure of the polar vortex and related changes in the stratospheric and tropospheric pressure fields and the stratospheric polar vortex affects surface weather patterns, the QBO has an effect on high latitude weather patterns (Baldwin and Dunkerton, 2001; Castanheira and Graf, 2003).



a



b

Fig. 10.3. Map (a) of magnitude of correlation between SST and the in-phase and quadrature 13.9 year SOI cycle. Dark grey (r = 0.7) and light grey (r = 0.6) and grey (r = 0.5) correspond to the area where the correlation coefficients are maximum. Map (b) of relative phase angle between the low-pass filtered (wavelength>10years) SST and in-phase and quadrature SOI 13.9 year cycle. Dark grey corresponds to the delay of about 2.5-3.0 years, light grey is area where the delay is about 1.5-2.0 year.

Jevrejeva et al. (2004) demonstrate that the delay of about two years in the 13.9 year signals detected in polar region can be explained by the transit time of the 13.9 year signal associated with equatorial coupled ocean (0.13-0.17 ms-1) propagation in the Pacific ocean (White, Tourre, Barlow and Dettinger, 2003), boundary Kelvin waves (1-3 ms-1) propagation along the western margins of the Americas (Meyers, Melsom, Mitchum and O'Brien, 1998) and by poleward-propagating of atmospheric angular momentum (Dickey, Marcus and Viron, 2003).

We have projected the 13.9 year signal extracted from SOI on the SST anomalies, which have been low-pass filtered (wavelength>10years) in order to remove typical 2.2-5.7 year ENSO signals SST (1854-1997) (Figure 10.3). The strength of correlation and the phase of correlation maps illustrate a meridional V-shape pattern in Pacific, which is symmetric along the equator, propagating eastward across the ocean and we can trace the propagation along the western margin of South and North America and also along the Arctic continental shelf as well as along Antarctica. The strongest effect is observed in polar regions.

5 Conclusion

We provide observational evidence of ENSO influence on the winter climate variability in NH during the last 135 years via signals in the 2.2, 3.5, 5.7 and 13.9 year bands. The presence of coupling between ENSO and NH ice conditions is revealed with use of wavelet coherence, average mutual information and mean phase coherence methods.

Phase relationships for the different frequency signals suggest that there are different mechanisms (fast and slow) for distribution of the 2.2-5.7 year and the 13.9 year signals. The 2.2- 5.7 year signals are most likely transmitted via the stratosphere, and the AO mediating propagation of the signals, through coupled stratospheric and tropospheric circulation variability that accounts for vertical planetary wave propagation. Low frequency variability (12-20 years) is more likely associated with an oceanic origin. The delay of about two years in the 13.9 year signals from Barents and Baltic Seas can be explained by long-lived ENSO events in the Pacific Ocean.

References

Baldwin, M.P. and Dunkerton, T.J. (2001) Stratospheric Harbingers of Anomalous Weather Regimes, *Science,* 294, 581-584.

Castanheira, J.M. and Craf, H.-F. (2003) North Pacific-North Atlantic relationships under stratospheric control?, *J. Geophys. Res*., 108, 4036, 10.1029/2002JD002754.

Dickey, J.O., S.L. Marcus and Viron, O. (2003) Coherent interannual and decadal variations in the atmosphere-ocean system, *Geophys. Res. Let,* 30, 2002GL016763.

Foufoula-Georgiou, E. and Kumar, K. (1995) *Wavelets in Geophysics*, Academic Press, 373.

Gloersen, R. 1995. Modulation of hemispheric sea-ice cover by ENSO events, *Nature*, 373, 503-505.

Grinsted, A., J. C. Moore and Jevrejeva, S. (2004) Application of the cross wavelet transform and wavelet coherence to geophysical time series, *Nonlinear Processes in Geophysics,* 11, 561-566.

Huang, J., K. Higuchi and Shabbar, A. (1998) The relationship between the North Atlantic Oscillation and the ENSO, *Geophys. Res. Let.*, 25, 2707-2710.

Jevrejeva, S. and Moore, J.C. (2001) Singular Spectrum Analysis of Baltic Sea ice conditions and large-scale atmospheric patterns since 1708, *Geophys. Res. Let.,* 28*,* 4503-07.

Jevrejeva, S., J. C. Moore and Grinsted, A. (2003) Influence of the arctic oscillation and El Nino-Southern Oscillation (ENSO) on ice conditions in the Baltic Sea: The wavelet approach, *J. Geophys. Res*., 108, 2003JD003417.

Jevrejeva, S., J.C. Moore and Grinsted, A. (2004) Oceanic and atmospheric transport of multiyear ENSO signatures to the polar regions. *Geophys. Res. Lett.*, 31, L24210, doi:10.1029/2004GL020871.

Kaplan, A., M. A. Cane, Y. Kushnir, A.C. Clement, M.B. Blumenthal and Rajagopalan, B. (1998) Analyses of global sea surface temperature 1856-1991, *J. Geophys. Res.,* 103, 18567-18589.

Merkel, U. and Latif, M. (2002) A high resolution AGCM study of the El Niño impact on the North Atlantic/European sector, *Geophys. Res. Let.,* 29, 2001GLO13726.

Meyers, S.D., A. Melsom, G.T. Mitchum and O'Brien, J.J. (1998) Detection of the fast Kelvin waves teleconnection due to El Niño Southern Oscillation, *J. Geophys. Res.*, 103, 27655-27663.

Mokhov, I. I. and Smirnov, D. A. (2006) El Niño–Southern Oscillation drives North Atlantic Oscillation as revealed with nonlinear techniques from climatic indices *Geophys. Res. Lett*., 33, L03708 10.1029/2005GL024557

Papoulis, A. (1984) *Probability, Random Variables, and Stochastic Processes*, second edition. New York: McGraw-Hill, *(See Chapter 15.)*

Pozo-Vázquez, D., M.J. Esteban-Parra, F.S. Rodrigo and Castro-Diez, Y. (2001) The association between ENSO and winter atmospheric circulation and temperature in the North Atlantic Region, *J. Clim.*, 14, 3408-3420.

Ribera, P., and Mann, M. (2002) Interannual variability in the NCEP reanalysis 1948-1999, *Geophys. Res. Let.*, 29, 2001GL013905.

Ropelewski, C.F., and Jones, P.D. (1987) An extension of the Tahiti-Darwin Southern Oscillation Index, *Monthly Weather Review, 115*, 2161-2165.

Seinä, A. and Palosuo, E. (1996) The classification of the maximum annual extent of ice cover in the Baltic Sea 1720-1995, *Report series of the Finnish Institute of Marine Research* No 27, 79-91.

Smith, T.M. and Reynolds, R.W. (2003). Extended Reconstruction of Global Sea Surface Temperatures Based on COADS Data (1854-1997), *J. Clim.*, 16, 1495-1510.

Thompson, D.W.J. and Wallace, J.M. (1998) The Arctic Oscillation signature in the winter geopotential height and temperature fields, *Geophys. Res. Let*., *25,* 1297-1300.

Torrence, C. and Compo, G. P. (1998) A practical guide to wavelet analysis, *Bull. Am. Meteorol. Soc.*, 79, 61–78.

Torrence, C. and Webster, P. (1999) Interdecadal Changes in the ENSO-Monsoon System, *J.Clim.,* 12, 2679-2690.

Venegas, S.A. and Mysak, L.A. (2000) Is there a dominant timescale of natural climate variability in the Arctic?, *J. Clim.,* 13, 3412-3434.

Venje, T. (2001) Anomalies and trends of sea ice extent and atmospheric circulation in the Nordic Seas during the period 1864-1998, *J. Clim.,* 14, 255-267.

White, W.B., Y.M. Tourre, M. Barlow, and Dettinger, M. (2003) A delayed action oscillator shared by biennial, interannual, and decadal signals in the Pacific Basin, *J. Geophys. Res.*, 108, 2002JC001490.

White, W.B. and Tourre, Y. M. (2003) Global SST/SLP waves during the 20th century, *Geophys. Res. Let*, 30, 2003JL017055.

11 25 Years of Nonlinearity in Oceanography from the Lagrangian Perspective

A. D. Kirwan, Jr.[1], B. L. Lipphardt, Jr.[2], A. C. Poje[3], L. Kantha[4] and M. Zweng[5]

[1] University of Delaware, College of Marine and Earth Studies, adk@udel.edu
[2] University of Delaware, College of Marine and Earth Studies, brucel@udel.edu
[3] City University of New York Staten Island, Dept. of Mathematics, poje@math.csi.cuny.edu
[4] University of Colorado, Department of Aerospace Engineering, kantha@colorado.edu
[5] University of Delaware, College of Marine and Earth Studies, mzweng@udel.edu

Abstract. The Lagrangian view of geophysical fluid dynamics relies heavily on nonlinear methods familiar to most geoscientists. These include identification of hyperbolic and elliptic regions in flow fields along with finite time and scale Lyapunov exponents and particle dispersion statistics. Here we identify elliptic and hyperbolic regions to study the life cycle of a large anticyclonic eddy in the Gulf of Mexico. Hyperbolic regions develop simultaneously down to 200m when the Loop Current sheds the ring. The ring's migration across the Gulf is monitored by tracking the movement of its elliptic point. Its abrupt disappearance is the result of interaction with a nearby hyperbolic region, which exists to 150m. Some broader implications of this approach are discussed in the last section.

1 Introduction

This conference has focused on the emergence of nonlinear methods in the geosciences during the past twenty-five years. This is an appropriate period for the rise of nonlinear technology in most subdisciplines in the geosciences. Euler established the central role of nonlinearity in fluid dynamics two and a half centuries ago. During most of the intervening 250 years the focus was on linear Eulerian dynamics, i.e. the equations of motion and conservation of mass were linearized. Solutions of this system of linear equations give the velocity as a function of position and time. The final solution step, neglected by most researchers during this period, is to calculate particle paths. Even if the dynamics that produce the velocity field is linear, the last step usually requires developing solutions to a system of nonlinear equations. The moral is that Lagrangian problems are inherently nonlinear, regardless of the Eulerian dynamics.

Saltzman (1962) first established the essential character of nonlinearity in atmospheric flow. He developed spectral solutions with time dependent amplitudes to the basic convection equations. Truncation of the series, even at low order, led to a system of nonlinear equations that did not display the regular character expected of the

perturbative methods in vogue at that time. Lorenz (1963) immediately realized the significance of Saltzman's results. In subsequent publications, he developed the theory and applied it to a variety of topics in atmospheric flow. Today, Lorenz is recognized as the founder of "chaos" or dynamical systems theory. This has emerged as the theoretical basis for investigations of nonlinear phenomena in all branches of science. It is appropriate for members of this conference to remember that atmospheric scientists started the modern theory of nonlinear science.

A quarter of a century ago Lagrangian geophysical fluid dynamics involved deploying acoustically tracked floats or satellite tracked drifters in the ocean and radio tracked balloons in the atmosphere. Considerable effort was spent trying to interpret the observed trajectories. For oceanographers, this was much like describing a bowl of spaghetti, hence the name "spaghetti diagrams." A recent example of such a plot, taken from Ohlman and Niiler (2005), is shown in Fig. 11.1. This figure summarizes the trajectories of over 750 drifters deployed along the Texas – Louisiana shelf from 1993 through 1998. There is a lot of information on ocean flow in the western Gulf of Mexico basin on time scales of days to months in this data, but it is not apparent in a cursory examination of the figure. Indeed the challenge to the authors was to cull the data to provide a cogent discussion of the statistical and dynamical properties of the circulation in this region of the Gulf.

Oceanographers now realize that the information content of trajectories varies greatly. In Fig. 11.1, for example, some of the trajectories provide mostly redundant information while a few others appear unique. In fact, knowledge of a few "critical" trajectories provides a template for assessing the origin and fate of many particles. This is actually an ocean application of a fundamental idea in chaos theory. There the concern is with identifying "basin boundaries of attraction", for open sets of initial conditions with different kinematic behavior. If the space is locally hyperbolic, then there are distinguished manifolds that intersect at unique "hyperbolic" trajectories. Since these manifolds consist of individual particle trajectories, they are themselves invariant under the flow and thus serve to as phase space partitions.

As pointed out by Aref (1984), in the context of fluid mechanical mixing and transport, the flow field is itself the phase space of interest. Consequently, the identification of manifolds provides a template for dividing extended sets of particle trajectories into regions with different origin and fate characteristics. In the context of standard analytic models of canonical geophysical flow fields, the mathematical theory is well developed in the two dimensional, time-periodic, perturbative case. Starting with a time-independent mean flow, one can completely quantify the particle transport across mean flow boundaries induced by the addition of a small, time-periodic velocity component. Once the manifolds are identified, the ideas of 'lobe-dynamics' (Ottino 1989, Rom-Kedar et al. 1990, Wiggins 1992) can be used to determine precisely which sets of particles enter and exit distinguished flow regions. The area or volume of these lobes along with the forcing frequency determines the advective flux of fluid between flow structures.

In the geophysical context, Samelson (1992) relied on these ideas in his study of the stirring properties of Gulf Stream meanders. Ridderinkhof and Loder (1994)

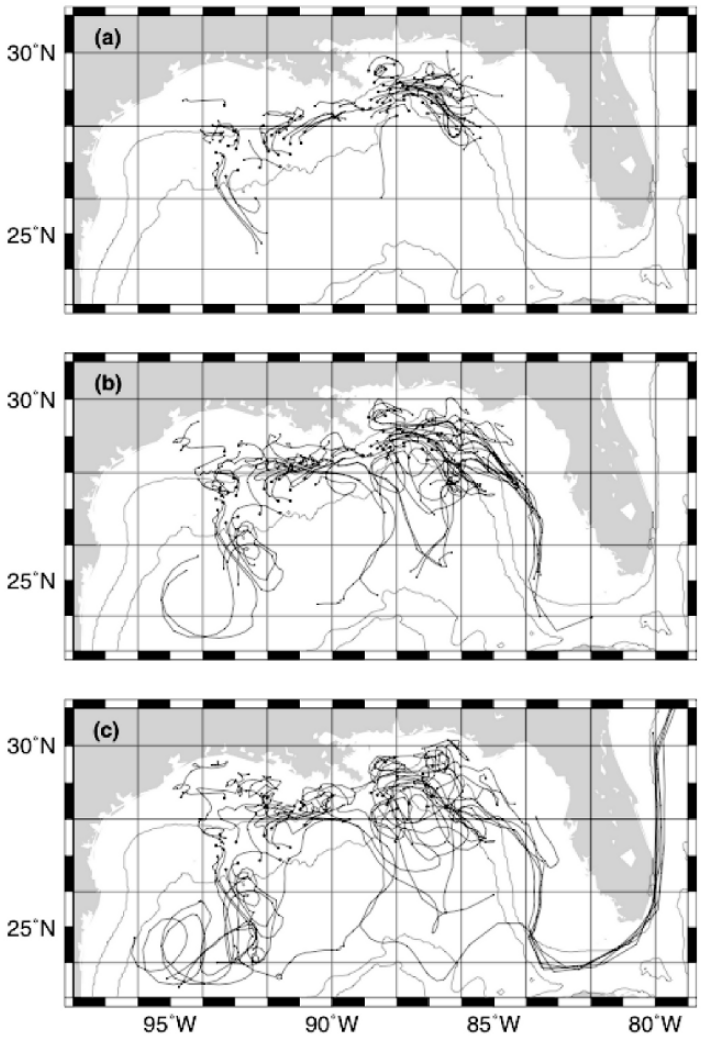


Fig. 11.1. Composite of drifter trajectories from the Gulf of Mexico. Taken from Ohlmann and Niiler (2005).

applied similar methods to tidally driven flow in a numerical model of the Gulf of Maine and Rogerson et al (1999) extended the Gulf Stream application to numerical solutions of dynamically consistent meandering jets.

A substantial challenge in recent years has been to further extend the notion of 'Lagrangian Coherent Structure' boundaries, classically defined by identifying invariant manifolds, to more complex and realistic geophysical velocity fields provided

either by data-assimilating numerical models or by high resolution quasi-synoptic data sets available from remote sensing instruments. The main mathematical issues involve confronting the necessarily finite-time, aperiodic nature of both the input Eulerian velocity and the Lagrangian transport phenomena of interest. For slowly evolving fields, Haller and Poje (1998) provide conditions for inferring the existence of finite-time manifolds in the Lagrangian frame from properties of frozen-time snapshots of saddle points in the Eulerian field. Alternately, local maxima in maps of finite-time or finite-size Lyapunov exponents computed from ensembles of trajectories have been shown to approximate transport boundaries in the Lagrangian frame (Haller 2000, Boffeta 2001, Haller 2002, Shadden et al. 2002). Computing maps of such quantities both forward and backward in time, locates controlling hyperbolic trajectories.

The methodology has, in general, matured to the point that it is commonly used in oceanographic applications. Using velocity fields from the data-assimilating numerical model of the Gulf of Mexico described here, Toner et al. (2003) were able to compute the manifolds delineating the material boundaries of a chlorophyll plume extending from the Yucatan to the West Florida Shelf. Earlier, Kuznetsov et al. (2002) examined in detail the Lagrangian kinematics of model ring pinch-off from the Loop Current. Transport barriers inferred from Lyapunov calculations have been used to shed light on the seasonality of mixing the Adriatic (d'Ovidio et al. 2004), pollution dispersal off Florida (Lekien et al. 2005), and the wide spatial variations in stirring properties of the Tasman Sea (Waugh, Abraham and Bowen 2006).

In geophysical fluid dynamics hyperbolic trajectories live in the deformation-dominated nether regions between eddies. In the ocean, these are near stagnant flow regions, but in the atmosphere, these regions may be advected by large-scale jet systems. Hyperbolic trajectories are one characteristic of chaos and low predictability, as explained by Kirwan et al. (2003). Elliptic trajectories are a less celebrated second type of critical trajectory. They are found near the centers of long-lived vortices where the flow is dominated by vorticity and predictability is high.

The purpose here is to provide a synopsis of some recent and promising developments in Lagrangian geophysical fluid dynamics. We hope that readers, who are active in nonlinear geoscience, but not particularly conversant with Lagrangian hydrodynamics, will benefit from the presentation. To this end, the paper is organized as follows. The next section provides an overview of the dynamics of both elliptic and hyperbolic trajectories and their relations to stagnation points. In Section 3, hyperbolic and elliptic trajectories are used to record the life cycle of a large eddy in the Gulf of Mexico. The final section discusses some broader implications of this work.

2 Critical points and trajectories

The traditional hyperbolic/elliptic characterization of fluid flows is based on the Okubo (1970) criterion, namely the eigenvalues of the horizontal Eulerian velocity gradient. Real (imaginary) eigenvalues occur in hyperbolic (elliptic) regions. In

singular regions of vanishing non-divergent velocity, real eigenvalues correspond to saddle points while imaginary eigenvalues correspond to centers. Maps of the Okubo-Weiss parameter given by the determinant of the deformation matrix have been used to partition oceanographic flow fields into shear and vorticity dominated regions.

These partitions are, however, inherently Eulerian quantities. Once time dependence is introduced, the connection between singular points in snapshots of the Eulerian field and critical particle trajectories is lost. The path of the evolving Eulerian 'stagnation' point need not be a solution of the Lagrangian equations and the existence of a persistent saddle (center) points in the Eulerian flow does not necessarily imply the existence of a hyperbolic (elliptic) trajectory in the particle dynamics. Haller and Poje (1998) studied the topology of the flow field in hyperbolic regions under the assumption of slow Eulerian evolution. They provide a set of sufficient, but not necessary, Eulerian conditions for the existence finite time hyperbolic trajectories in resulting Lagrangian flow.

In short, Haller and Poje show that if there exists a sufficiently strong saddle point in the Eulerian flow field, persisting for a sufficiently long enough time, there must exist a nearby hyperbolic particle trajectory with associated stable and unstable finite time manifolds. The directions of these manifolds at their intersection are closely aligned with the eigendirections of the Eulerian velocity gradient. The manifold lying along the positive (negative) eigendirection is called the unstable (stable) or outflowing (inflowing) manifold. The hyperbolic trajectory acts as a local magnet for the flow field. Particles near the inflowing manifold are attracted exponentially in time toward it while particles near the outflowing manifold are repelled exponentially in time.

As such, assuming the sufficient conditions are satisfied one can attempt to locate hyperbolic trajectories by searching for saddle type stagnation points in the Eulerian flow field. This is important for analysis since stagnation points (local velocity minima), which are easily observed in Eulerian flows, can serve as an initial guess for the hyperbolic trajectory location. Thus, their study provides the theoretical basis for many computational and observational geophysical fluid dynamics investigations. Since their criterion does not care about the dynamic processes that produce a given velocity field it has wide applicability.

The theory does not guarantee that every stagnation point has an associated hyperbolic trajectory, and the explicit existence conditions are often too involved for practical implementation in oceanographic data sets. This motivated Kirwan (2006) to develop a local f-plane diagnostic model of hyperbolic and elliptic trajectories. The latter are elliptic region counterparts of hyperbolic trajectories. As this model is dynamic, there are four evolution equations, two Euler equations for the velocity, and two path equations relating the velocity field to trajectories. The resulting frequency equation is quartic rather than quadratic, as in the Okubo case. Hence, there are two pairs of eigenvalues. In hyperbolic regions, one pair of eigenvalues is real and the other is imaginary. The real eigenvalues determine the manifolds while the imaginary pair acts to impart anticyclonic spirals to the manifolds as they approach

or leave the hyperbolic trajectory. Being dynamic, this model also provides criteria as to how near the hyperbolic trajectory and stagnation point are to each other. However, this is not sufficient to locate either hyperbolic or elliptic trajectories. Consequently, users still employ 'straddling' techniques based on seeding persistent observed Eulerian saddle points. We are not aware of any procedure that locates elliptic trajectories that may be near the relative stagnation points in eddies.

This model gives a physically satisfying picture of stagnation points and associated critical trajectories. If the flow is steady then the two coincide. But, once time dependency in the geopotential field is introduced, stagnation points and critical trajectories separate. The reason is that the critical trajectories are Lagrangian objects and thus must obey the equations of motion, while stagnation points are Eulerian objects whose paths are not subject to dynamical constraints. Also, if geostrophic dynamics are assumed then the model reduces to the Okubo case.

The next section uses these concepts to describe the formation and unexpected breakup of a large anticyclonic ring in the Gulf of Mexico. The elliptic stagnation point is used to track the ring from birth to its demise in the middle of the Gulf. The birth of the ring is denoted by the development of a hyperbolic trajectory at the separatrix between the ring and the Loop Current. The ring breakup is caused by the strong deformation associated with an inflowing manifold from a nearby hyperbolic trajectory.

3 Birth and death of Eddy Fourchon in the Gulf of Mexico

3.1 Background

Since the pioneering work of Elliott (1982), it is widely recognized that large anticyclonic rings shed by the Loop Current dominate the circulation in the Gulf of Mexico. The Loop Current is a component of a continuous Atlantic Ocean western boundary current, which enters the Gulf through the Straits of Yucatan and egresses to the Atlantic through the Florida Straits. The prevailing view is that these rings migrate across the Gulf and slowly spin down along the western slope approximately one year after being shed. At birth, these rings are among the largest in the world ocean. Diameters are of the order of 200 to 300 km with swirl speeds exceeding 2 m/s. An excellent summary of the Gulf of Mexico circulation and ring characteristics is available in Sturges and Lugo-Fernandez (2005).

Figure 11.2, taken from Lipphardt et al. (2006) shows the mean velocity and eddy kinetic energy at 50 meters for 12 November 1997 through 30 September 2001. The Loop Current dominates both the eddy kinetic energy and the mean currents in the eastern Gulf. Apparently, the Loop Current was well extended into the Gulf during a significant portion of the analysis period. The western Gulf mean current field indicates a well-defined western boundary current as the western limb of a large anticyclonic circulation. Kantha et al. (2005) reported that six rings were shed from the

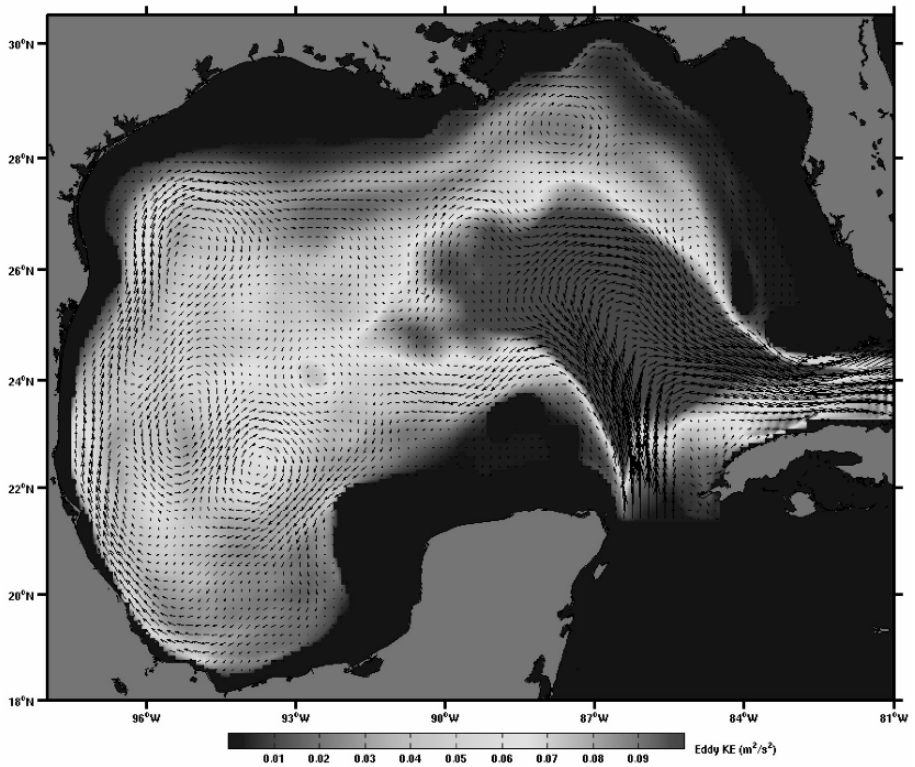


Fig. 11.2. Color contours of eddy kinetic energy (m^2/s^2) and mean velocities at 50 m from the University of Colorado Princeton Ocean Model. Taken from Lipphardt et al. (2006). (A color version of this figure appears between pages 196 and 197).

Loop Current during this period. These rings made little impact on the eddy kinetic energy statistics. Presumably, they were obscured by the extended Loop Current in the eastern basin. There is no indication in the eddy kinetic energy of rings migrating to the western slope as theory predicts. Instead, there is a local maximum near 93^0 W.

3.2 Birth of eddy Fourchon

Eddies and rings in the mid-ocean are well documented by hydrographic and altimeter data. The same data leave no question as to the origins of these vortices. However, documenting eddy or ring pinchoff events from jets is a longstanding problem for oceanographers. Hydrographic data provide detailed information on the vertical structure of the ocean, but horizontal resolution is poor and, because of the time required to complete ship surveys, it does not provide a synoptic picture during pinchoffs. Altimeter data can provide excellent information about near surface

conditions but little information on the vertical structure. Moreover, these data are not synoptic. Data assimilating numerical models are quite useful for blending hydrographic and altimetric data into synoptic pictures that have been used to document ring formation in the Gulf of Mexico and elsewhere. See Kantha et al. (2005) for an excellent discussion of the technical issues and examples.

Typically, these studies attempt to characterize eddy formation by closed contours of streamline-like quantities, such as geopotential height or altimetric height anomalies. These models and the basic data have produced metrics that characterize eddy separation. Despite these developments, there often is uncertainty as to eddy formation. Sometimes formation metrics indicate that an eddy has separated but later reattaches. These metrics are useful for describing characteristics of mature eddies, but it is not clear that they provide reliable pictures during the periods of rapid evolution, such as occurs during a pinchoff. The fundamental issue is that streamlines are not material lines when the flow is unsteady.

A new metric is proposed here. When an eddy or ring separates there should be a material surface that bars advective transport between the vortex and the parent jet. In the simplest picture of eddy formation, a jet meander becomes so extreme it reattaches. When this happens, a stagnation point develops suggesting there is an associated hyperbolic trajectory. We illustrate this criterion by reexamining the formation of Eddy Fourchon; one of the six eddies formed during the analysis period covered by Fig. 11.2.

Using a data assimilating hindcast model Kantha et al. (2005) analyzed 16 pinchoff events over a nine-year period. Their criteria for pinchoff were a combination of closed sea surface height contours and model velocities at 50m. They reported the formation date for Fourchon as 19 February 1998. As shown in Fig. 11.3, a hyperbolic trajectory first develops at 50m on 15 February and lasts until 26 February. The inflowing manifold is shown as the blue curve and the outflowing manifold as the red curve in this figure. The red circles show the local elliptic points for Fourchon and an adjacent cyclone to the southwest. Note that the hyperbolic trajectory develops in the nether region between these two vortices, a weak cyclonic circulation to the northeast, and the anticyclonic circulation associated with the Loop Current. Both manifolds wrap around Fourchon and clearly show that the anticyclone is no longer exchanging fluid with the Loop Current. The inflowing manifold wraps around the eastern limb of Fourchon while the outflowing manifold wraps around the western limb.

We located hyperbolic trajectories and calculated manifolds for every model level from the surface to 300m. The times of the first appearance of hyperbolic trajectories and the character of the manifolds are consistent with Fig. 113 to 300m. The situation at 300m is different as seen in Fig. 11.4. The hyperbolic trajectory appears on 19 February, the same date given by Kantha et al. (2005). Moreover, both manifolds now wrap around the cyclone to the southwest. However, the elliptic points for Fourchon and the cyclone are still present.

This analysis suggests that for this ring, the pinchoff occurred simultaneously on 15 February 1998 at each of the model levels down to a level of 300m. The manifolds

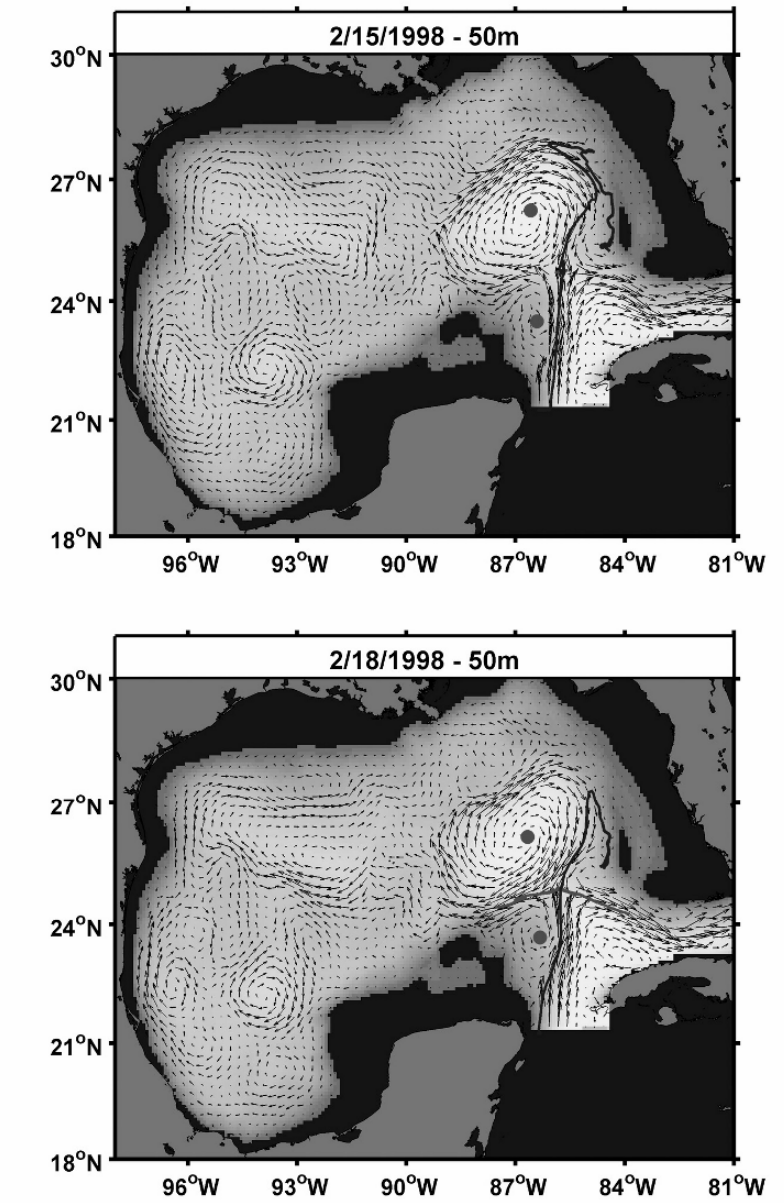


Fig. 11.3a. Manifolds and hyperbolic trajectory at 50 m at the beginning of the Fourchon pinchoff (15 and 18 February 1998). The inflowing (outflowing) manifolds are blue (red) and the elliptic points for Fourchon and an adjacent cyclone are red circles. (A color version of this figure appears between pages 196 and 197).

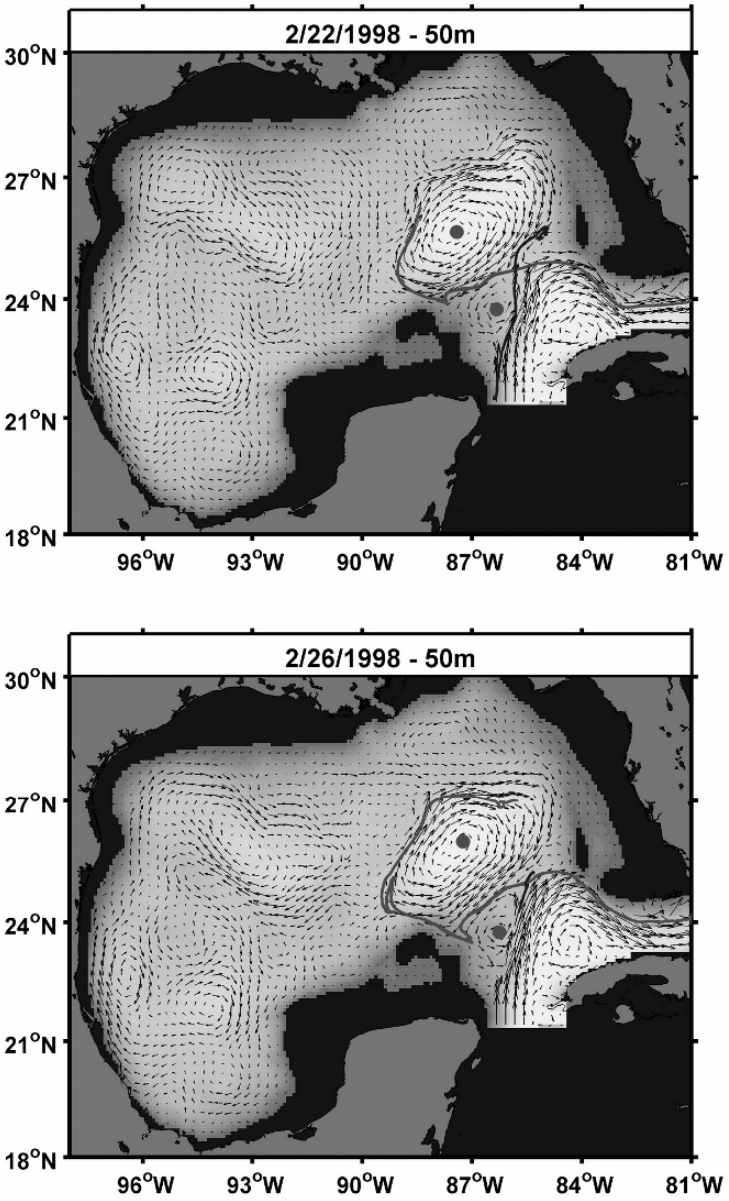


Fig. 11.3b. Manifolds and hyperbolic trajectory at 50 m at the end of the Fourchon pinchoff (22 and 26 February 1998). The inflowing (outflowing) manifolds are blue (red) and the elliptic points for Fourchon and an adjacent cyclone are red circles. (A color version of this figure appears between pages 196 and 197).

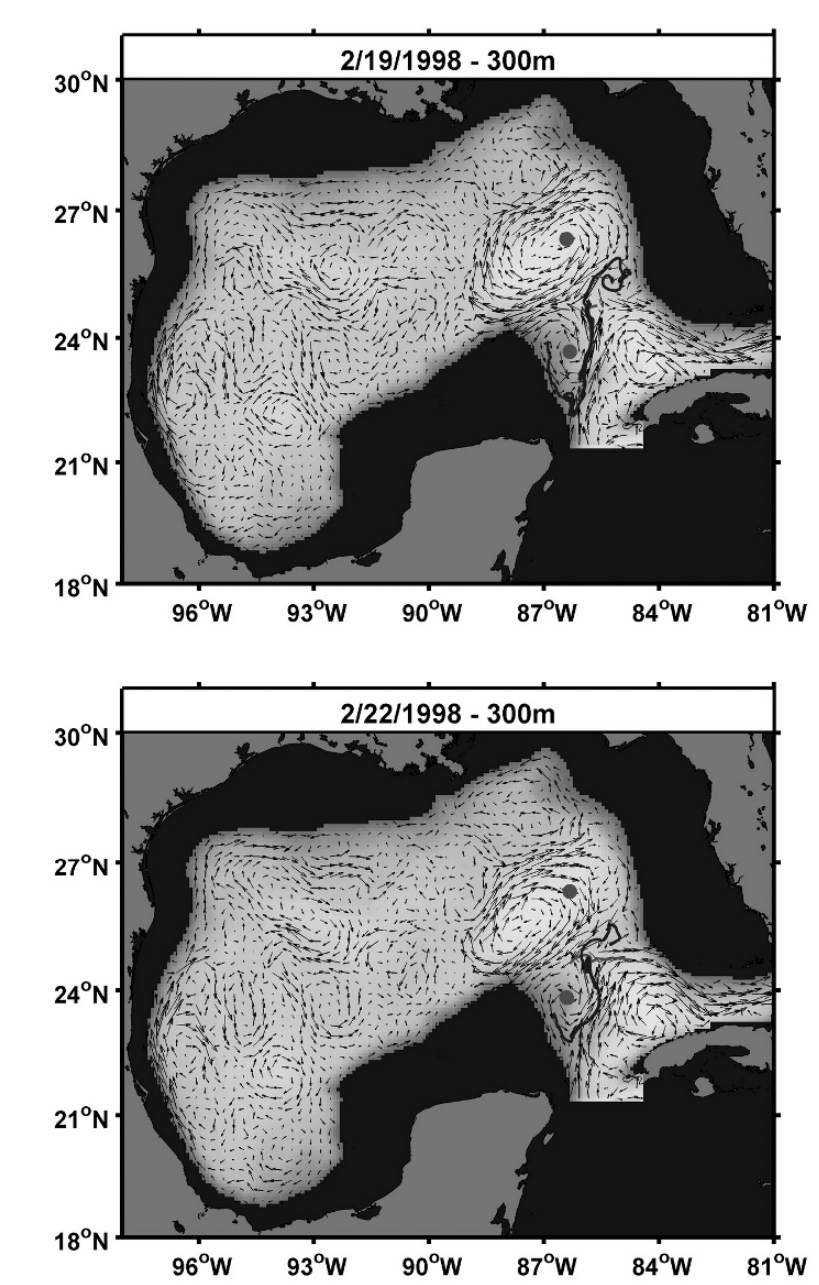


Fig. 11.4a. Similar to Fig. 11.3a but at a depth of 300 m for 19 and 22 February 1998. (A color version of this figure appears between pages 196 and 197).

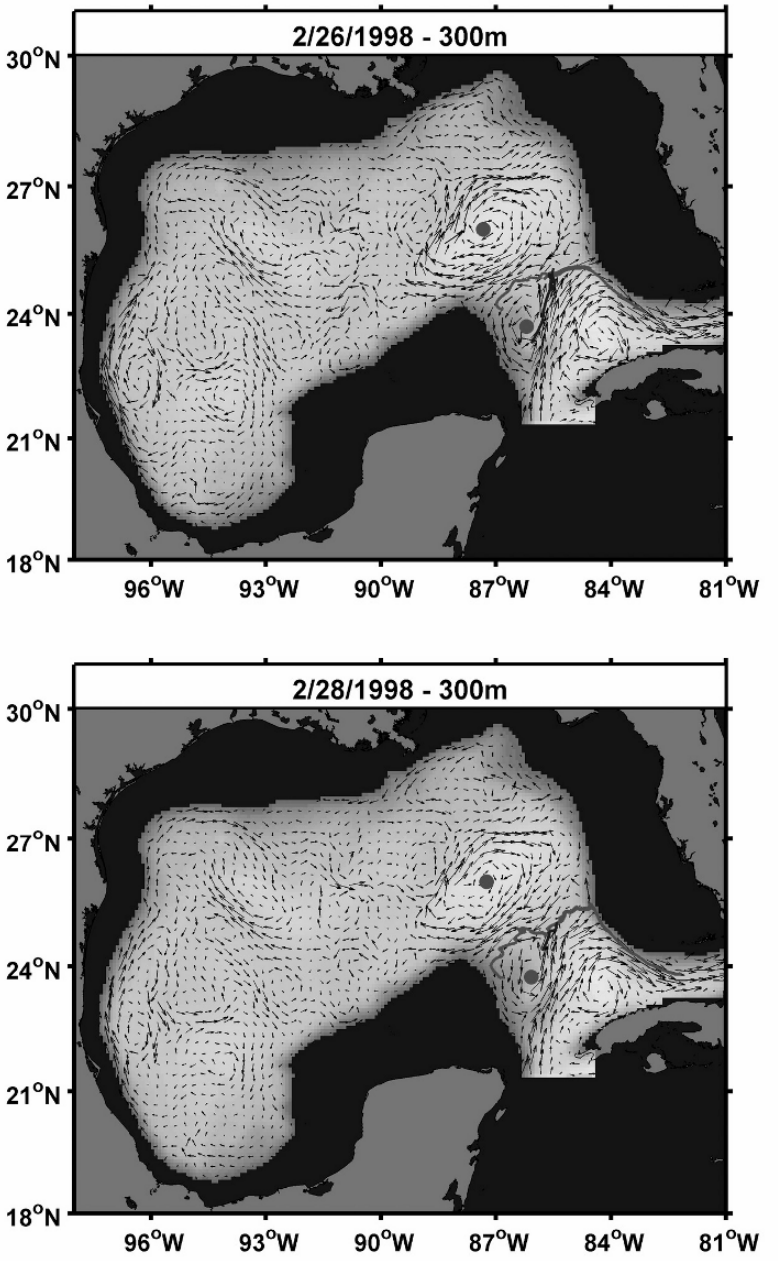


Fig. 11.4b. Similar to Fig. 11.3b, but at a depth of 300m for 26 and 28 February 1998. (A color version of this figure appears between pages 196 and 197).

at this latter level occurred at a slightly later date and do not seem to be connected to the pinchoff process at the shallower levels.

3.3 Short life of eddy Fourchon

A variety of methods have been used to track rings. Kantha et al. (2005) and Lipphardt et al. (2006) have summarized the attributes and deficiencies of many of these methods. Here we follow Lipphardt et al. (2006) and track Fourchon by tagging the elliptic point associated with its local velocity minimum.

Figure 11.5 shows the path of the elliptic points for 50 and 150m. We were unable to track elliptic points below the latter depth shortly after its birth. This is not surprising since rings are lens shaped and tend to shallow as they age. This figure shows that the paths at the two depths nearly coincide. Neither elliptic point could be tracked beyond 6 August 1998. This finding agrees with the analysis of Kantha et al. (2005). The next subsection addresses the question of why this ring could not be tracked further.

3.4 Death of eddy Fourchon

The inability to track a vigorous ring such as Fourchon for more than 6 months in a high resolution, data assimilating model is puzzling. We have considered three possibilities. The simplest explanation is that the metrics used to track it are not sufficiently precise to detect it beyond 92°W. This explanation cannot be discounted. Although different metrics produced the same result, they both rely on the same basic data. The second possibility is that the altimeter data is too coarse to track this feature. Resolution of either of these possibilities will require considerable testing of assimilation algorithms along with carefully planned field experiments.

The third possibility is simpler: the ring breaks up abruptly around 92°W. What mechanisms could produce such a traumatic event? Lipphardt et al. (2006) found a hyperbolic trajectory that developed just to the west of Fourchon on 17 July 1998. One limb of the inflowing manifold cuts through the middle of Fourchon. As this manifold winds up into the hyperbolic trajectory, it cuts the ring in half. Figures 11.6 and 11.7 show the results at 50 and 150 meters. The figures are completely consistent. They show projections of a two dimensional material surface at these two levels that was generated by a baroclinic hyperbolic structure.

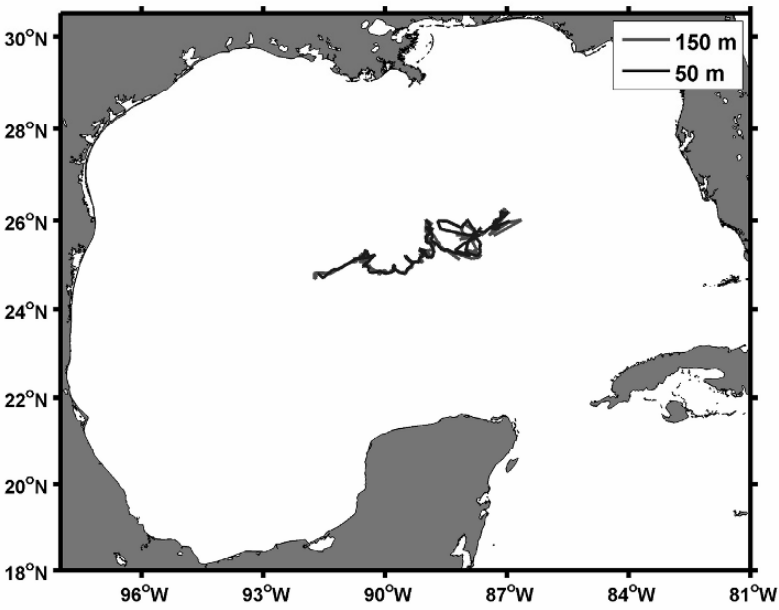


Fig. 11.5. Paths of the elliptic points of Fourchon at 50 and 150 m from formation to breakup in the middle of the Gulf. (A color version of this figure appears between pages 196 and 197).

4 Discussion

The analysis presented here applied elementary Lagrangian methods to the output of a data-assimilating numerical model of the Gulf of Mexico. The focus was on identification of critical points and trajectories in the flow field. The notion of hyperbolicity is central to chaos theory, which is the theoretical basis for much of nonlinear analysis in the geosciences. Since the flow field is the phase space in geophysical fluid dynamics, the results of such analysis are readily apparent.

In this study the formation of hyperbolic trajectories and associated material manifolds gave a slightly earlier birth date for a Loop Current ring than that provided by conventional Eulerian methods. It is interesting that hyperbolic trajectories developed simultaneously to 300m during formation.

Presently, there is no technology that determines elliptic trajectories that may live near the elliptic points at the centers of vortices. Consequently, we used the elliptic point itself to track Eddy Fourchon. This point was not tracked west of 92°W; a finding completely consistent with standard Eulerian methods used to track eddies. Further

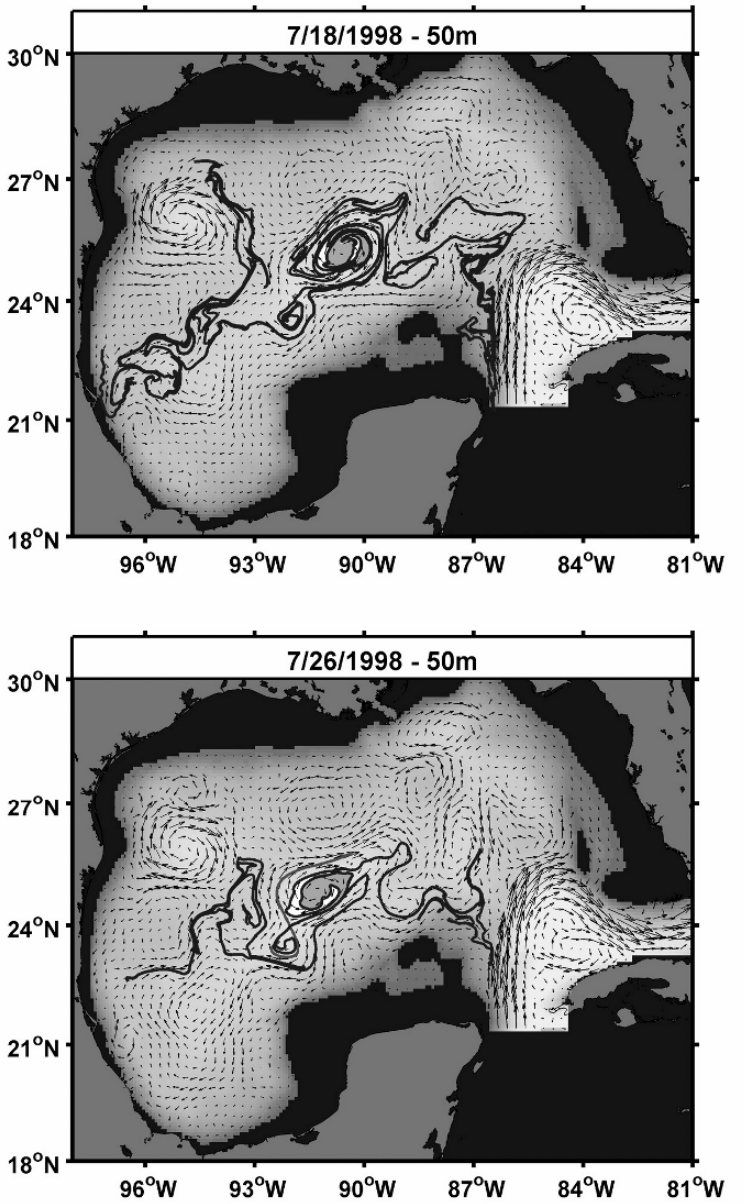


Fig. 11.6a. Manifolds and hyperbolic trajectory at 50 m at the start of the Fourchon breakup (18 and 26 July 1998). The parts of the eddy that are on different sides of the inflowing manifold are colored differently for emphasis. (A color version of this figure appears between pages 196 and 197).

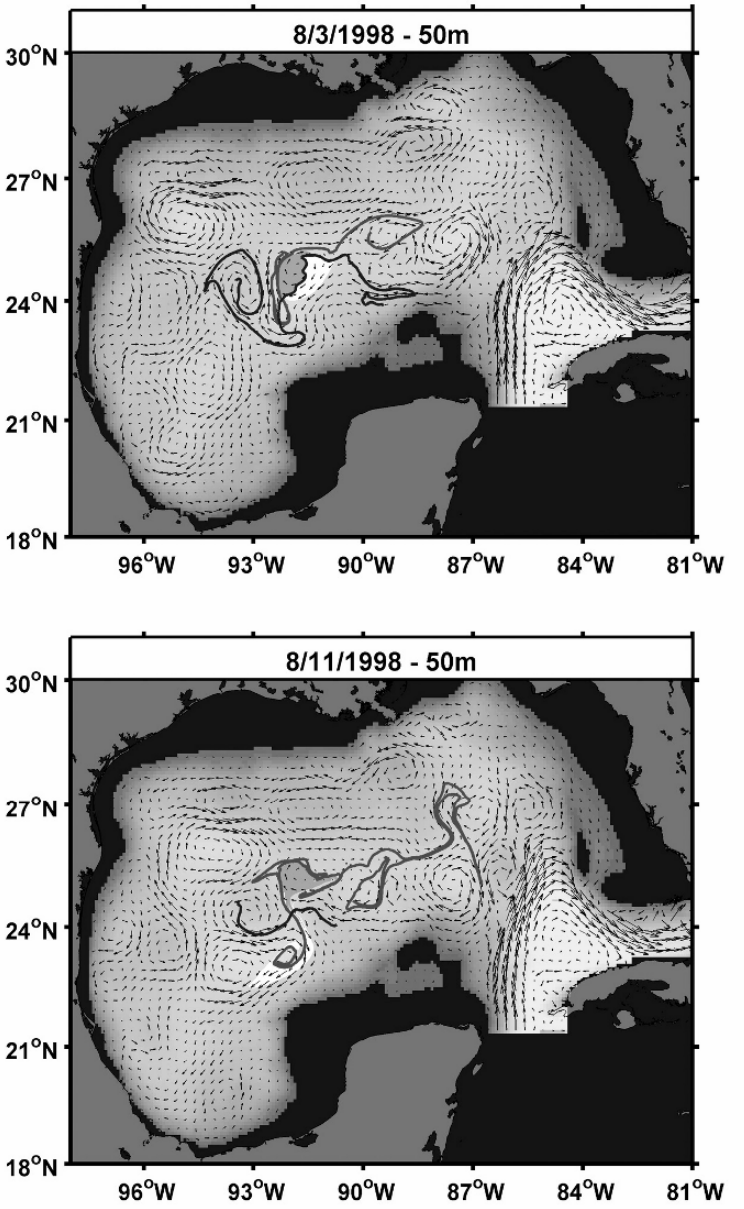


Fig. 11.6b. Manifolds and hyperbolic trajectory at 50 m at the end of the Fourchon breakup (3 and 11 August 1998). The parts of the eddy that are on different sides of the inflowing manifold are colored differently for emphasis. (A color version of this figure appears between pages 196 and 197).

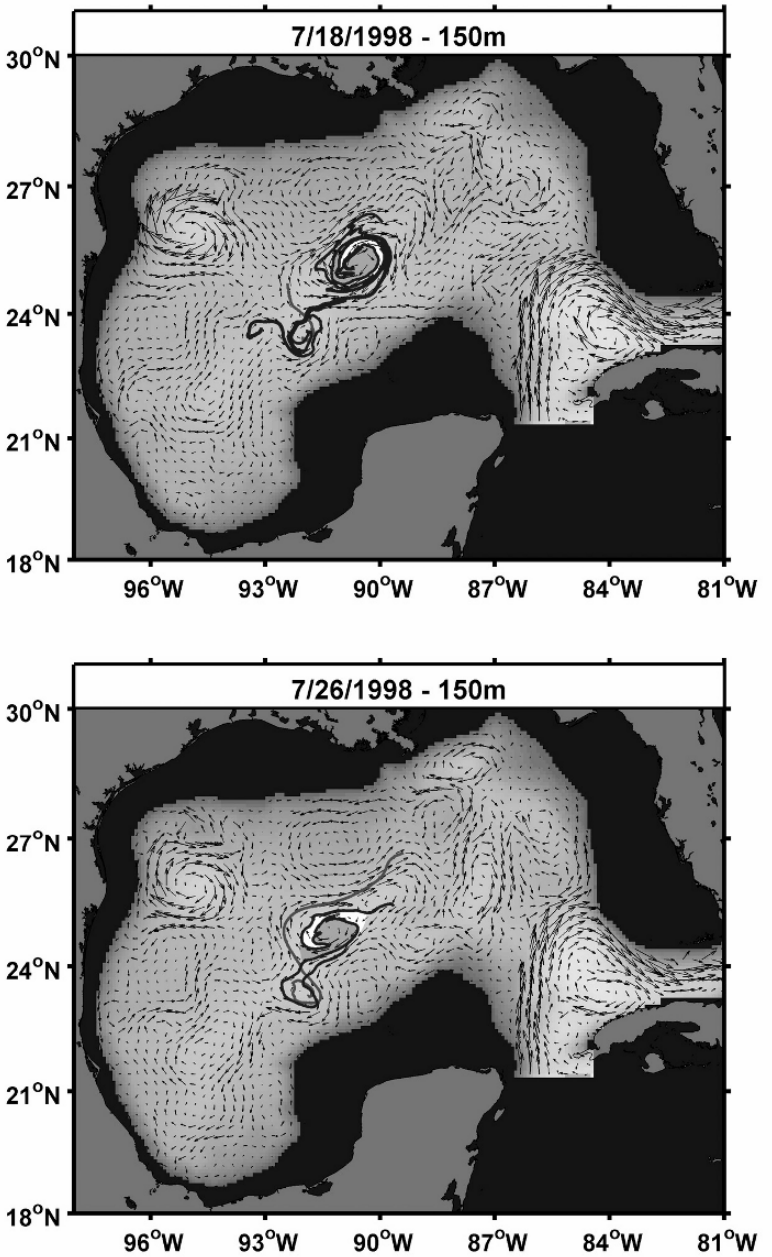


Fig. 11.7a. Same as Fig. 11.6a, except at 150 m. (A color version of this figure appears between pages 196 and 197).

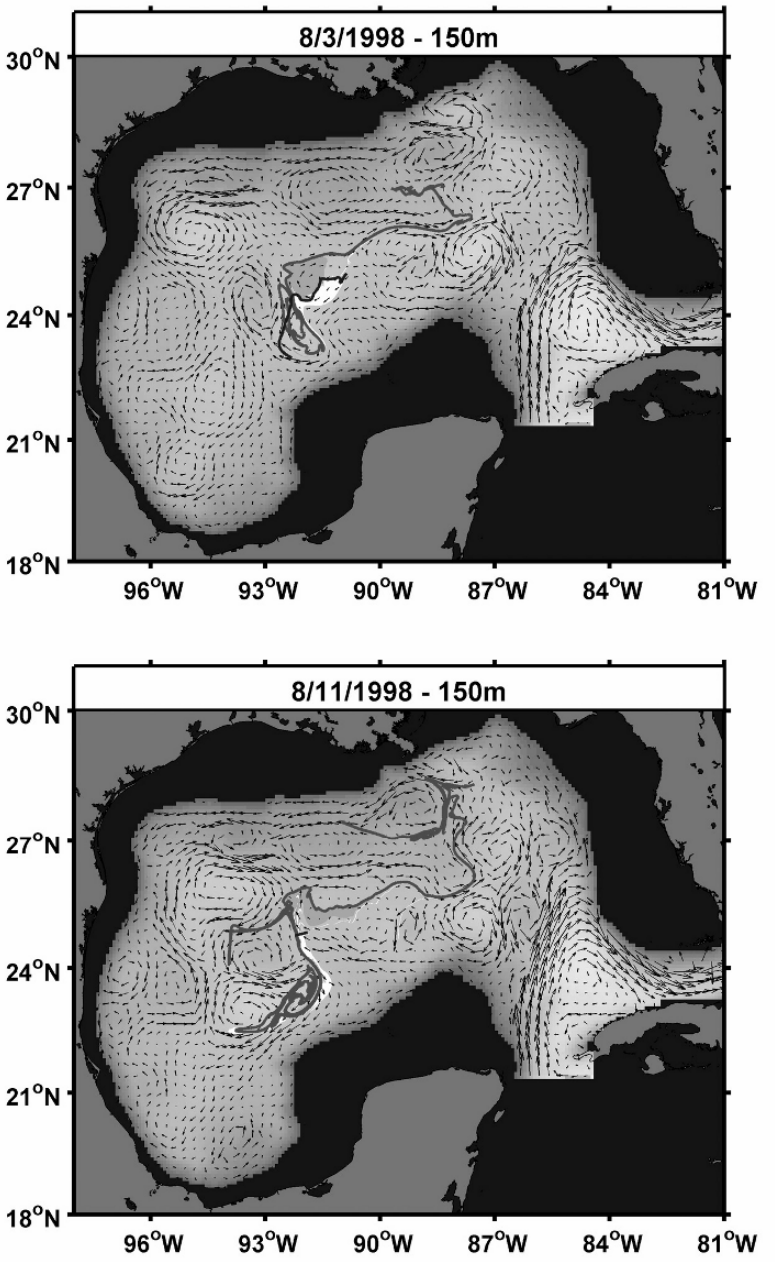


Fig. 11.7b. Same as Fig. 11.6b, except at 150 m. (A color version of this figure appears between pages 196 and 197).

analysis showed that the ring was abruptly torn apart by the inflowing manifold associated with a nearby hyperbolic trajectory. This manifold was tracked to 150m so the breakup process occurred simultaneously down this level.

The abrupt breakup of this ring is surprising since it is widely believed that these large rings spin down much further to the west, along the continental slope. Moreover, there is ample evidence that ocean eddies are long-lived. In this regard, it is stressed that the breakup is not an artifact of the Lagrangian methods. These methods merely showed that the breakup was caused by nearby baroclinic hyperbolic structures. The real issue is whether these structures are the artifact of assimilation of poorly resolved altimeter data into a highly resolved model, or whether the model was able to correctly infer this feature from noisy data. Obviously, much more research is required before this issue can be settled.

Nevertheless, this result is consistent with the Kantha et al. (2005) analysis of 16 rings over a nine-year period. Using Eulerian methods, they claimed that only one ring could be tracked to the western slope. Lipphardt et al. (2006) reanalyzed that ring, using a wide variety of Lagrangian methods, and found that it also broke up around 92°W. The picture of the western Gulf that emerged from their study is that it is a region of ephemeral writhing mesoscale eddies. Transient hyperbolic regions develop as these eddies interact and the associated manifolds act to both destroy and create eddies in this region.

This report has been concerned with only a few Lagrangian methods currently in use. For example, Lipphardt et al. (2006) augmented these methods with particle dispersion and finite-scale Lyapunov exponent calculations that revealed mesoscale phenomena not reported by Eulerian methods. This demonstrates that Lagrangian methods are a useful supplement to traditional approaches for model assessment as they reveal flow details not easily extracted from Eulerian based analyses.

Acknowledgments

C. Blett provided invaluable technical assistance. Support for this research was provided by grants N00014-00-0019 and N00014-05-1-0092 from the Office of Naval Research and grants 0222150 and ATM-200102 from the National Science Foundation and through grants to the University of Delaware, and the Mary A. S. Lighthipe endowment.

References

Boffetta, G., Lacorata, G., Radaelli, G., and Vulpiani, A. (2001). Detecting barriers to transport: A review of different techniques. Physica D, 159, 58-70.

d'Ovidio, F., Fernandez, V., Hernandez-Garcia, E., and Lopez, C. (2004). Mixing structures in the Mediterranean Sea from finite-size Lyapunov exponents. GRL, 31, L17203.

Elliott, B. A. (1982). Anticyclonic rings in the Gulf of Mexico. Journal of Physical Oceanography, 12, 1292-1309.

Haller, G. and Poje, A. C. (1998). Finite time transport in aperiodic flows. Physica D, 119, 352-380.

Haller, G. (2000). Finding finite-time invariant manifolds in two-dimensional velocity fields. Chaos, 10, 99-108.

Haller, G. (2002). Lagrangian coherent structures from approximate velocity data. Phys. Fluids, 14, 1851-1861

Kantha, L., Choi, J.-K., Schaudt, K. J., and Cooper, C. K. (2005). A regional data assimilating model for operational use in the Gulf of Mexico. In: Sturges, W. and Lugo-Fernandez, A. (editors) (2005). Circulation in the Gulf of Mexico: Observations and Models. American Geophysical Union Press, 165-180.

Kirwan, A. D., Jr., Toner, M. and Kantha, L. (2003). Predictability, uncertainty, and hyperbolicity in the ocean. International Journal of Engineering Science, 41, 249-258.

Kirwan, A. D., Jr. (2006). Dynamics of "critical" trajectories. Progress in Oceanography, August-September, 448-465, doi: 10.1016/j.pocean.2005.07.002.

Kuznetsov, L., Toner, M., Kirwan, A. D., Jones, C., Kantha, L., and Choi, J. (2002). The Loop Current and adjacent rings delineated by Lagrangian analysis. J. Mar. Res. 60.

Lipphardt, B. L., Jr., Poje, A. C., Kirwan, A. D., Jr., Kantha, L., and Zweng, M. (2006). Death of three Loop Current rings. Submitted, Journal of Marine Research.

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. Journal of the Atmospheric Sciences, 20, 130-141.

Okubo, A. (1970). Horizontal dispersion of floatable particles in the vicinity of velocity singularities such as convergences. Deep-Sea Research, 17, 445-454.

Olhmann, J. C. and Niiler, P. P. (2005). Circulation over the continental shelf in the northern Gulf of Mexico. Progress in Oceanography, 64, 45-81.

Ottino, J. (1989). The Kinematics of Mixing: Stretching, Chaos, and Transport. Cambridge University Press, Cambridge, MA.

Ridderinkhof, H., and Zimmerman, J. (1992). Chaotic stirring in a tidal system. Science, 258, 1107-1111.

Rogerson, A., Miller, P., Pratt, L., and Jones, C.K.R.T. (1999). Lagrangian motion and fluid exchange in a barotropic meandering jet. J. Phys. Oceanogr. 29, 2635–2655.

Rom-Kedar, V., Leonard, A., and Wiggins, S. (1990). An analytical study of transport, mixing and chaos in an unsteady vortical flow. J. Fluid Mech. 214, 347–394.

Saltzman, B. (1962). Finite Amplitude Free Convection as an Initial Value Problem – I. Journal of the Atmospheric Sciences, 19, 329-341.

Samelson, R. M. (1992). Fluid exchange across a meandering jet. Journal of Physical Oceanography, 22, 431-440.

Shadden, S. C., Lekien, F., and Marsden, J. E. (2005). Definition and properties of Lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows. Physica D, 212, 271–304.

Sturges, W. and Lugo-Fernandez, A. (editors) (2005). Circulation in the Gulf of Mexico: Observations and Models. American Geophysical Union Press.

Toner, M., Kirwan, A. D., Jr., Poje, A. C., Kantha, L. H., Müller-Karger, F. E., and Jones, C. K. R. T. (2003). Chlorophyll dispersal by eddy-eddy interactions in the Gulf of Mexico. Journal of Geophysical Research, 108, 3105, doi:1029/2002JC001499.

Waugh, D. W., Abraham, E. R., and Bowen, M. M. (2006). Spatial variations of stirring in the surface ocean: A case study of the Tasman Sea. Journal of Physical Oceanography, 36, 526–542.

Wiggins, S. (2005). The Dynamical Systems Approach to Lagrangian Transport in Oceanic Flows. Annual Reviews of Fluid Mechanics, 37, 295-328, doi:10.1146/annurev.fluid. 37.061903.175815.

12 Self-Scaling of the Statistical Properties of a Minimal Model of the Atmospheric Circulation

Valerio Lucarini[1,2], Antonio Speranza[1] and Renato Vitolo[1]

[1] University of Camerino, Department of Mathematics and Computer Science, PASEF, Camerino, Italy, lucarini@alum.mit.edu

[2] University of Bologna, Department of Physics, ADGB – Atmospheric Dynamics Group Bologna, Bologna, Italy

Abstract. A quasi-geostrophic intermediate complexity model of the mid-latitude atmospheric circulation is considered, featuring simplified baroclinic conversion and barotropic convergence processes. The model undergoes baroclinic forcing towards a given latitudinal temperature profile controlled by the forced equator-to-pole temperature difference T_E. As T_E increases, a transition takes place from a stationary regime - Hadley equilibrium - to a periodic regime, and eventually to a chaotic regime, where evolution takes place on a strange attractor. The dependence of the attractor dimension, metric entropy, and bounding box volume in phase space is studied by varying T_E. It is found that this dependence is smooth and has the form of a power-law scaling. The observed smooth dependence of the system's statistical properties on the external parameter T_E is coherent with the chaotic hypothesis proposed by Gallavotti and Cohen, which entails an effective structural stability for the attractor of the system. Power-law scalings with respect to T_E are also detected for global observables responding to global physical balances, like the total energy of the system and the averaged zonal wind. The scaling laws are conjectured to be associated with the statistical process of baroclinic adjustment, decreasing the equator-to-pole temperature difference. The observed self-similarity could be helpful in setting up a theory for the overall statistical properties of the general circulation of the atmosphere and in guiding - also on a heuristic basis - both data analysis and realistic simulations, going beyond the unsatisfactory mean field theories and brute force approaches.

1 Introduction

The problem of understanding climate is an issue of scientific as well as of practical interest, since it is connected to aspects of vital importance for human life and environment. Recently, climate has also become a politically relevant issue. In this chapter we propose our scientific approach and some encouraging results concerning the theory of General Atmospheric Circulation (GAC) - the core machine of weather and climate - looked upon as a problem in Physics. Climate is defined as the set of

the statistical properties of the climatic system. In its most complete definition, the climatic system is composed of four intimately interconnected sub-systems, atmosphere, hydrosphere, cryosphere, and biosphere. These subsystems interact nonlinearly with each other on various time-space scales (Lucarini 2002). The atmosphere is the most rapid component of the climatic system, features both many degrees of freedom, which makes it complicated, and nonlinear interactions of different components involving a vast range of time scales, which makes it complex. The GAC is a realization of planetary-scale thermodynamic transformations in a rotating, stratified fluid. Among all the physical processes involved in the GAC, the baroclinic conversion plays a central role because it is through this mechanism that the available potential energy (Margules 1903; Lorenz 1955), stored in the form of thermal fluctuations, is converted into the vorticity and kinetic energy of the air flows. Apart from the classical role of baroclinic instability (Charney 1947; Eady 1949) in mid-latitude weather development (Dell'Aquila et al. 2005), different forms of baroclinic conversion operate at larger scales, *e.g.* in the ultra-long planetary waves associated with the so called low-frequency variability (Charney and DeVore 1979; Benzi et al. 1986; Benzi and Speranza 1989; Ruti et al. 2006), and at smaller scales, *e.g.* in the banded sub-frontal structures (Mantovani and Speranza 2002).

Historically - see the classical monograph by Lorenz (1967) - the problem of GAC has been essentially approached in terms of analyzing the time-mean circulation. Usual separations such as between time average and fluctuations, or between the zonal and the eddy field, are somewhat arbitrary, but perfectly justifiable with arguments of symmetry, evaluation of statistical moments, etc.. The time average - fluctuations separation is often suggestive of a theory in which the fluctuations growing on an unstable basic state, identified with the time average, feed back onto the basic state itself and *stabilize* it. The central objective of the classical circulation theory is the *closure* in the form of a *parameterization* of the nonlinear eddy fluxes in terms of quantities which can be derived from the mean field itself. This approach is not only unsuccessful, it is just not feasible, because the stability properties of the time mean state do not provide even a zeroth-order approximation of the dynamical properties of the full nonlinear system. This has been illustrated by theoretical arguments (Farmer 1982) and counter-examples of physical significance (Speranza and Malguzzi 1988; Malguzzi et al. 1990). As a consequence, even sophisticated versions of the idea that the GAC dynamics may be defined in terms of the statistics derivable from average atmospheric fields have proved unfruitful for guiding the interpretation of observations and the development of models.

Alternative approaches have been proposed, *e.g.* mimicking the GAC via dynamical systems approach with simplified low-dimensional models, say below 10 degrees of freedom (Ghil et al. 1985). However, the physical relevance and mathematical generality of these approaches have been criticized as well, whereas the relevance of the space-time geophysical scaling behavior has been emphasized (Scherzer and Lovejoy 2004). Nevertheless, a comprehensive and coherent picture of the GAC is still far from being available and the classical *time-mean* approach has not been completely abandoned as a paradigm.

It is clear that in the context of the classic paradigm of GAC, the main task of modeling consists in capturing the mean field: the fluctuations will follow as its instabilities. This point of view, whether explicit or not, has strongly influenced the set-up and diagnostics of existing General Circulation Models (GCMs). As a consequence, the diagnostic studies usually focus on comparing the temporal averages of the simulated fields rather than on analyzing the dynamical processes provided by the models. Similar issues are related to the provision of the so-called extended range weather forecasts. Suppose, in fact, that the forecaster was given the next month average atmospheric fields: what practical information could he derive from that? Of course, if dynamical information were stored in the average fields - typically in the form of dominant regimes of instability derivable from the time-mean flow - he could obtain useful information from the prediction of such time mean fields. Actually, the problem of extended range forecast is still open even in terms of clearly formulating what we should forecast in place of average fields.

Profound scientific difficulties arise when considering the GAC from a statistical mechanics point of view. In spite of several attempts (see e.g. Leith 1975), analogues of the fluctuation-dissipation theorem (Kubo 1966), or of Kramers-Kronig relations (Lucarini et al. 2005) cannot be straightforwarly formulated: the dynamical features of a forced and dissipative chaotic system are completely different from those of a Hamiltonian system immersed in a thermal bath. Therefore, the problem of constructing a Climate Change theory on this basis is rather delicate, essentially because the GAC features the non-equivalence between the external and internal fluctuations (Lorenz 1979). The main reason for this non-equivalence is that, because of dissipation, the attractor of the system lives on a manifold of zero volume inside its phase space: the internal natural fluctuations occur within such a manifold whereas externally induced fluctuations would move the system out of the attractor with probability 1. In other words, ergodicity does not hold in the whole phase space, as usually assumed in equilibrium statistical mechanics along the Boltzmann paradigm. In this case, ergodicity - if present at all - is restricted to the attractor, which is identified as a Sinai-Ruelle- Bowen (SRB) measure (Eckmann and Ruelle 1985). Recently Ruelle (1998) has derived a generalization of the fluctuation-dissipation theorem in forced and dissipative systems, under the assumption that a unique SRB measure exists. This concept also lies at the basis of the Chaotic hypothesis formulated by Gallavotti and Cohen (1996) and Cohen and Gallavotti (1999). The project of writing sets of equations directly describing the statistics of the system on the attractor – the projection of the original equations of motion onto the attractor - has been carried on with great attention in the recent past (Lorenz 1980; Fritz and Robinson 2001). But, despite the big efforts, no applicable result has been obtained so far (Foias and Olson 1996).

In the absence of robust scientific paradigms, a gap was created between the studies revolving around phenomenological and/or numerical modeling issues and those focused on fundamental mechanisms of GAC (Held 2005). The assumption that adopting models of ever increasing resolution will eventually lead to the final understanding of the GAC (a sort of *brute force* approach) is not based on any

consolidated mathematical and physical knowledge and may, in fact, result misleading. One basic reason is that, in the limit of infinite resolution for any numerical model of fluid flow, the convergence to the statistical properties (infinite time) of the continuum real fluid flow dynamics is not guaranteed. As a matter of fact, we note that at the present stage, most of the leading climate models are neither consistent nor realistic even in the representation of the basic time-space spectral properties of the variability of the atmosphere at mid-latitudes (Lucarini et al. 2006).

Finally, problems arise when auditing climate models, since we have always to deal with three qualitatively distinct attractors: the attractor of the real atmosphere; the reconstruction of the real attractor from observations; the attractors of the model (maybe more than one) we adopt in climatic studies. The first one is obviously unknown; as for the second one, the (presumably) best reconstructions available are provided by the reanalyses (Kalnay 2003). Note that presently there are two alternative main global reanalyses, which feature good but not perfect agreement in their statistical properties (Dell'Aquila et al. 2005; Ruti et al. 2006). As for the third kind of attractors, the various GCMs surely possess substantially different attractors, which reflects into disagreements in the statistical properties of the atmosphere (Lucarini et al. 2006).

This chapter wishes to gives hints for addressing the following question: once we abandon the unsatisfactory mean field theories, what may come next? What may lead a new approach to a theory of GAC, which can serve as a guidance for setting up and diagnosing GCMs? Which statistical properties may be crucial? Therefore, we are, at this stage, not seeking realism in the model representation of atmospheric dynamics, but rather searching for representations of key nonlinear processes with the minimum of ingredients necessary to identify the properties under investigation. We have decided to choose a rather simplified quasi-geostrophic (QG) model with just a few hundreds degrees of freedom, able to capture the central processes of the mid-latitude portion of the GAC. In Sec. 2 we discuss the derivation of the evolution equations for the two-layer QG model. In Sec. 3 we characterize the transition from stationary to chaotic dynamics and study the dependence on T_E and on model resolution JT of the dimension of the strange attractor, of the metric entropy, and of the volume of its bounding box in the phase space. In Sec. 4 we analyze the statistics of the total energy of the system and the latitudinally averaged zonal wind. In Sec. 5 we give our conclusive remarks and perspectives for future works.

2 The Model

The description of the large scale behavior of the atmosphere is usually based on the systematic use of dominant balances, which are derived on a phenomenological diagnostic basis, but whose full-correctness at theoretical level is still unclear. When considering the dynamics of the atmosphere at mid-latitudes, on spatial and temporal scales comparable with or larger than those of the synoptic weather (about 1000 Km and 1 day, respectively), the hydrostatic and geostrophic balances are phenolmenologically well-established. From the set of *ab-initio* dynamic and thermodynamic

equations of the atmosphere it is possible to obtain a set of simplified prognostic equations for the synoptic weather atmospheric fields in a domain centered at mid-latitudes - the QG equations - by assuming that the fluid obeys the hydrostatic balance and undergoes small departures from the geostrophic balance (Charney 1948; Pedlosky 1987; Speranza and Lucarini 2005). A great number of physical phenomena are *filtered out* of the equations by the QG approximation: various types of waves associated with strong local divergence, turbulent motions, etc.. There is no doubt that these are small on the time-space scales of the motions we consider, but it is still an open question to what extent they influence or not the statistics of large scale atmospheric motions and, in case, how to model such a statistical effect. Note that in general the QG attractor is not a good approximation to the attractor of the corresponding full *ab-initio* equations, despite the fact that the QG balance approximation is diagnostically quite good for the dominating time-space scales of the atmosphere (Lorenz 1980).

In this work we consider a *β-channel* where the zonal coordinate x is periodic with period L_x and the latitudinal coordinate $y \in [0, L_y]$. As a further approximation, we consider only two vertical layers (Phillips 1954). This is the minimal system retaining the baroclinic conversion process. In order to avoid problems in the definition of the boundary conditions, due to the prescription of the interaction with the polar and the equatorial circulations (Speranza and Malguzzi 1988), we consider a domain extending from the pole to the equator. Near the equator the QG approximation is not valid, so that the representation of the tropical circulation is beyond the scope. The equations of motion are:

$$D_t^1\left(\Delta_H \psi_1 + f_0 + \beta y\right) - f_0 \frac{\omega_2 - \omega_0}{\delta p} = 0, \tag{1}$$

$$D_t^3\left(\Delta_H \psi_3 + f_0 + \beta y\right) - f_0 \frac{\omega_4 - \omega_2}{\delta p} = 0, \tag{2}$$

$$D_t^2\left(\frac{\psi_1 - \psi_3}{\delta p}\right) - H_2^2 \frac{f_0}{p_2^2}\omega_2 = \kappa \Delta_H \left(\frac{\psi_1 - \psi_3}{\delta p}\right) + \frac{R}{p_2 f_0}\frac{Q_2}{C_p}. \tag{3}$$

where f_0 is the Coriolis parameter and β its meridional derivative evaluated at the center of the channel, κ parameterizes the heat diffusion, R and C_p are the thermodynamical constants for dry air, Δ_H is the horizontal Laplacian operator. The streamfunction ψ_j is defined at pressure levels $p = p_{j=1} = p_0/4$ and $p = p_{j=3} = 3/4\ p_0$, while the vertical velocity ω is defined at $p = p_{j=0} = 0$ (top boundary), $p = p_{j=2} = p_0/2$, and $p = p_{j=4} = p_0$ (surface boundary). The pressure level pertaining to the approximation to vertical derivative $\partial\psi_g/\partial p$ as well as the stratification height H is $p = p_{j=2}$, and $\delta p = p_3 - p_1 = p_0/2$. The Lagrangian derivative defined at the pressure level p_j is written as $D_t^j \bullet = \partial_t + J(\psi_j, \bullet)$, where J is the Jacobian operator defined as $J(A,B) = \partial_x A \partial_y B - \partial_y A \partial_x B$. The streamfunction at the level p_2 is computed as an

average between the streamfunctions of the levels p_1 and p_3, so that the Lagrangian derivative at p_2 is $D_t^2 = 1/2\left(D_t^1 + D_t^3\right)$. The diabatic heating Q_2 is defined as:

$$Q_2 = \nu_N C_p\left(T^* - T_2\right) = \nu_N C_p \frac{f_0 p_2}{R}\left(\frac{2\tau^*}{\delta p} - \frac{\psi_1 - \psi_3}{\delta p}\right), \tag{4}$$

where the temperature T_2 is defined at p_2 and is defined via hydrostatic relation. This implies that the system is relaxed towards a prescribed temperature profile T^* with a characteristic time scale of $1/\nu_N$. T^* and τ^* are defined as follows:

$$T^* = \frac{T_E}{2}\cos\left(\frac{\pi y}{L_y}\right), \quad \tau^* = \frac{R}{f_0}\frac{T_E}{4}\cos\left(\frac{\pi y}{L_y}\right), \tag{5}$$

so that T_E is the forced temperature difference between the low and the high latitude boundaries of the domain. We assume no time dependence for T_E, thus discarding the seasonal effects. Since by thermal wind relation $\left(\vec{u}_1 - \vec{u}_3\right)/\delta p = \hat{k} \times \vec{\nabla}/\left(\psi_1 - \psi_3\right)/\delta p$, we have that the diabatic forcing Q_2 induced a relaxation of the vertical gradient of the zonal wind $u_1 - u_3$ towards the prescribed profile $2m^* = 2\mathrm{d}\tau^*/\mathrm{d}y$.

By imposing $\omega_0 = 0$ (top of the atmosphere), and by assuming $\omega_4 = -E_0\Delta_H\psi_3$ Ekman pumping (Pedlosky 1987) - from Eqs. (1)-(3) one obtains the evolution equations for the barotropic $\phi = 1/2\left(\psi_1 + \psi_3\right)$ and baroclinic field $\tau = 1/2\left(\psi_1 - \psi_3\right)$:

$$\partial_t \Delta_H \phi + J\left(\phi, \Delta_H \phi + \beta y\right) + J\left(\tau, \Delta_H \tau\right) = -\frac{2\nu_E}{H_2^2}\Delta_H\left(\phi - \tau\right). \tag{6}$$

$$\begin{aligned} \partial_t \Delta_H \tau - \frac{2}{H_2^2}\partial_t \tau + J\left(\tau, \Delta_H \phi + \beta y + \frac{2}{H_2^2}\phi\right) + J\left(\phi, \Delta_H \tau\right) = \\ \frac{2\nu_E}{H_2^2}\Delta_H\left(\phi - \tau\right) - \frac{2\kappa}{H_2^2}\Delta_H \tau + \frac{2\nu_N}{H_2^2}\left(\tau - \tau^*\right), \end{aligned} \tag{7}$$

where $\nu_E = f_0 E_0 H_2^2/\left(2\delta p\right)$ is the viscous-like coupling between the free atmosphere and the planetary boundary layer via Ekman pumping.

The two-layer QG system can be brought to non-dimensional form, which is more usual in the meteorological literature and is easily implementable in computer codes. This is achieved by introducing length and velocity scales l and u and performing a non-dimensionalization of both the system variables $\left(x, y, t, \phi, \tau, T\right)$ (as described in Table 12.1) and of the system constants (Table 12.2). In our case appropriate values are $l = 10m^6$ and $u = 10ms^{-1}$.

In this work we consider a simplified spectral version of Eqs. (6)-(7), where truncation is performed in the zonal Fourier components so that only the zonally symmetric component and one of the non-symmetric components are retained. The derivation is reported in Speranza and Malguzzi (1988). The main reason for this choice is that we wish to focus on the interaction between the zonal wind and waves,

Table 12.1. Variables of the system and their scaling factors.

Variable	Scaling Factor	Value of the scaling factor
x, y	L	$10^6\, m$
t	$u^{-1}l$	$10^5\, s$
$\psi_1, \psi_3, \phi, \tau$	Ul	$10^7\, m^2 s^{-1}$
U, m	U	$10\, ms^{-1}$
$\lambda_1, \ldots, \lambda_{6JT}$	ul^{-1}	$10^{-5}\, s^{-1}$
t_p	$u^{-1}l$	$10^5\, s$
T	$ul\, f_0 R^{-1}$	$3.5\, K$
E	$u^2 l^2$	$5.1 \times 10^{17}\, J$

thus neglecting the wave-wave nonlinear interactions. Since quadratic nonlinearities generate terms with Fourier components corresponding to the sum and difference of the Fourier components of the two factors, we can exclude direct wave-wave interactions provided that we only retain a single wave component (see, *e.g.*, Lucarini et al. (2005) for a general discussion of these effects in a different context). Note that if cubic nonlinearities were present, direct self wave-wave interaction would have been possible (Malguzzi and Speranza 1981; Benzi et al 1986). In the present case, the wave can self-interact only indirectly through the changes in the values of the zonally symmetric fields. This amounts to building up equations which are *almost-linear*, in the sense that the wave dynamics is linear with respect to the zonally symmetric parts of the fields (*i.e.*, the winds). The wavelength of the only retained wave component is $L_x / 6$, since we intend to represent the baroclinic conversion processes, which in the real atmosphere take place on scales of $L_x / 6$ or smaller (Dell'Aquila at al. 2005). In so doing we are retaining only one of the classical ingredients of GAC, *i.e.* the *zonal wind-wave* interaction.

Each of ϕ and τ are thus determined by 3 real fields: the zonally symmetric parts and the real and imaginary part of the only retained zonal Fourier component. A pseudospectral decomposition with *JT* modes is then applied to the resulting 6 real

Table 12.2. Parameters of the system and their scaling factors. The system is equivalent to what presented in Malguzzi and Speranza (1988) with the following correspondences: $1/H_2^2 \leftrightarrow F$, $\nu_N \leftrightarrow \nu_H$, $2\nu_E/H_2^2 \leftrightarrow 2\nu_E$, $2\kappa/H_2^2 \leftrightarrow \nu_S$.

Parameter	Dimensional Value	Non-dimensional value	Scaling factor
L_x	$2.9 \times 10^7\, m$	29	l
L_y	$10^7\, m$	10	l
H_2	$7.07 \times 10^5\, m$	7.07×10^{-1}	l
f_0	$10^{-4}\ s^{-1}$	10	ul^{-1}
β	$1.6 \times 10^{-11}\, m^{-1} s^{-1}$	1.6	ul^2
ν_E	$5.5 \times 105\, m^2 s^{-1}$	5.5×10^{-2}	ul
ν_N	$1.1 \times 10^{-6}\, s^{-1}$	1.1×10^{-1}	ul^{-1}
κ	$2.75 \times 10\, m^2 s^{-1}$	2.75×10^{-1}	ul
T_E	28 K to 385 K	8 to 110	$ul\, f_0 R^{-1}$

fields in the y -direction, yielding a set of $6 \times JT$ ordinary differential equations in the spectral coefficients. For the truncation order JT we have used the values $JT = 8$, 16, 32, 64. This is the prototypal model we use as *laboratory* for the analysis of the GAC. The only form of realism we are trying to achieve here is that the statistical properties of the nonlinear cycle *baroclinic instability - barotropic & baroclinic stabilization* are represented.

3 Dynamical and Statistical Characterization of the Model Attractor

The system of equations (6)-(7) has a stationary solution describing a zonally symmetric circulation, characterized by the stationary balance between the horizontal temperature gradient and the vertical wind shear. This solution corresponds on the Earth system to the idealized pattern of the Hadley equilibrium (Held and Hou 1980). There is a value of the equator-to-pole temperature gradient T_E^H such that if $T_E < T_E^H$ the Hadley equilibrium is stable and has a virtually infinite basin of attraction, whereas if $T_E > T_E^H$ it is unstable. In the stable regime with $T_E < T_E^H$, after the decay of transients, the fields ϕ, τ, and T are time-independent and feature zonal symmetry - they only depend on the variable y. Moreover, they are proportional by the same near-to-unity factor to the corresponding relaxation profiles. In particular, this implies that all the equilibrium fields are proportional to the parameter T_E.

When increasing the values of the control parameter T_E beyond T_E^H, the equilibrium becomes unstable - as first pointed out on vertically continuous models (Charney 1947; Eady 1949) and on the two two-layer model (Phillips 1954) - with respect to the process of baroclinic conversion. This allows for the transfer of available potential energy of the zonal flow stored into the meridional temperature gradient into energy of the eddies, essentially transferring energy from the latter component to the first two components of the energy density expression, later presented in Eq. (10). The observed value of T_E^H increases with the considered truncation order JT (see Table 12.3). A finer resolution allows for more efficient stabilizing mechanisms, because they act preferentially on the small scales. Such mechanisms are the barotropic stabilization of the jet, increasing the horizontal shear through the convergence of zonal momentum, which is proportional to the quadrature of the spatial derivatives of the fields ϕ and τ (Kuo 1973), and the

Table 12.3. Values of T_E defining the loss of stability of the Hadley equilibrium (T_E^H) and the onset of the chaotic regime (T_E^{crit}).

JT	T_E^H	T_E^{crit}
8	7.83	9.15
16	8.08	8.42
32	8.28	8.52
64	8.51	8.66

viscous dissipation, which is proportional to the Laplacian of the fields ϕ and τ. This is a clarifying example that, in principle, it is necessary to include suitable renormalizations in the parameters of a model when changing the resolution *JT*, in order to keep correspondence with the resulting dynamics (Lorenz 1980). In our case the values of T_E^H obtained for the adopted resolutions are rather similar. A stable periodic orbit branches off from the Hadley equilibrium as T_E increases above T_E^H. The attracting periodic orbit persists for T_E in a narrow interval, where it disappears through a saddle-node bifurcation taking place on an attracting invariant two-torus. For slightly larger values of T_E, *i.e.* for $T_E > T_E^{crit}$, a strange attractor develops by so-called quasi-periodic breakdown of a *doubled* torus. This is one of the most typical routes for onset of chaos (weak turbulence) in fluid dynamics experiments and low-dimensional models, compare Broer et al. (2002) and Randriamampianina et al. (2005) and references therein. The invariant objects involved in the transition to low-dynamical chaos correspond to well-known fluid flow patterns. In particular, the two-torus attractor in phase space yields an *amplitude vacillation* in the flow (Randriamampianina et al. 2005). For *JT* = 8, a different route to chaos takes place, involving a quasi-periodic Hopf bifurcation of the two-torus (instead of a quasi-periodic period doubling), whereby an invariant three-torus is created. The question remains open whether the signature of the bifurcations of the system, of the kind exploited in the context of physical oceanography in a recent book (Dijkstra 2005), is present in the geometrical structure of the fully developed strange attractor and is potentially useful for computing the statistical properties.

3.1 Lyapunov exponents and dimension of the strange attractor

In order to characterize the dynamical properties of the strange attractors of the truncated system, we study of the Lyapunov exponents (Oseledec 1968; Eckmann and Ruelle 1985), which are denoted, as customary, by $\lambda_1, \lambda_2, ..., \lambda_N$, with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$, $N = 6 \times JT$. The maximal exponent λ_1 becomes positive as T_E crosses the torus breakdown value T_E^{crit}, and then increases monotonically with T_E. Despite the schematic approach adopted in this work, the obtained Lyapunov spectra are qualitatively similar to what reported in Vannitsem (1997) and Snyder and Hamill (2003), where a much larger number of degrees of freedom was considered. The distribution of the exponents approaches a smooth shape for large T_E, which suggests the existence of a well-defined *infinite baroclinicity* model obtained from as a - possibly, singular perturbation - limit for $T_E \to \infty$. We will analyze elsewhere this mathematical property, which is possibly of physical interest as well.

The Lyapunov exponents are used to compute the Lyapunov dimension (also called Kaplan-Yorke dimension, see Eckmann and Ruelle (1985) and Kaplan and Yorke (1979) and metric entropy - also known as Kolmogorov-Sinai entropy (Eckmann and Ruelle 1985). The Lyapunov dimension is defined as $D_L = k + \sum_{i=1}^{k} \lambda_i / |\lambda_{k+1}|$, where k is the unique index such that $\sum_{i=1}^{k} \lambda_i \geq 0$ and $\sum_{i=1}^{k+1} \lambda_i < 0$. Under general assumptions on the dynamical system under examination, D_L is an upper bound for the Hausdorff dimension of an attractor. The estimates for

the dimension of an attractor obtained by computing the correlation and information dimensions (Farmer et al. 1983) become completely meaningless when the Lyapunov dimension increases beyond, say, 20. In particular, the correlation and information algorithms tend to drastically underestimate the dimension. This is a well-known problem: for large dimensions, prohibitively long time series have to be used. Ruelle (1990) suggests as a rule of thumb that a time series of 10^d statistically independent data are needed in order to estimate an attractor of dimension d. Therefore, computational time and memory constraints in fact limit the applicability of correlation-like algorithms only to low-dimensional attractors.

The number of positive Lyapunov exponents - unstable dimension (Eckmann and Ruelle 1985) - increases with T_E, which implies that the Lyapunov dimension also does so. This is confirmed by a plot of the Lyapunov dimension as a function of T_E shown in Fig. 12.1 left. For small values of $(T_E - T_E^{crit})$, we have that $D_L \propto (T_E - T_E^{crit})^{\gamma}$, with γ ranging from ≈ 0.5 ($JT = 8$) to ≈ 0.7 ($JT = 64$). The range of T_E where this behavior can be detected increases with JT. For larger values of T_E a linear scaling regime of $D_L \approx \eta T_E + const.$ is found in all cases. The linear coefficient is for all JT remarkably close to $\eta \approx 1.2$. The domain of validity of the linear approximation is apparently homothetic - see the simple geometric construction in Fig. 12.1 left. For T_E larger than a JT-depending threshold, there occurs a sort of phase-space saturation, as the Lyapunov dimension begins to increase sublinearly with T_E. Note that while for $JT = 8$ the model is in this regime in most of the explored T_E - domain ($T_E \geq 20$), for $JT = 64$ the threshold is reached only for $T_E \approx 108$. In this latter regime of parametric dependence the system is not able to provide an adequate representation of the details of the dynamics of the system.

The metric entropy $h(\rho)$ of an ergodic invariant measure ρ expresses the mean rate of information creation, see Eckmann and Ruelle (1985) for definition and other properties. If a dynamical system possesses a SRB invariant measure ρ , then $h(\rho)$ is equal to the sum of the positive Lyapunov exponents. Existence of an SRB measure is hard to prove for a given nonhyperbolic attractor. It has been only proven

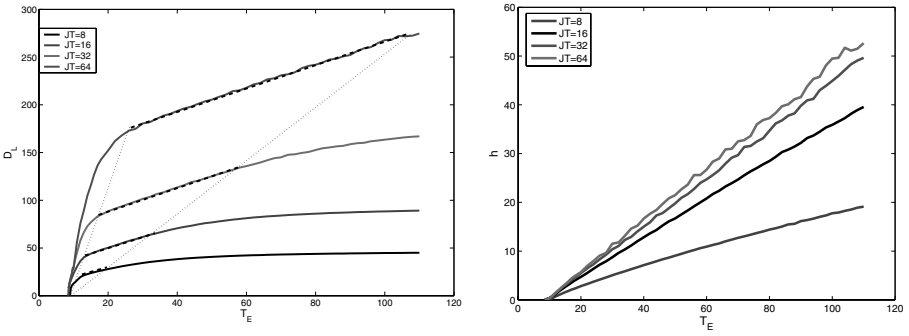


Fig. 12.1. Left: Lyapunov dimension of the attractor of the model as a function of T_E. All the straight lines are parallel and the domain of validity of the linear fit is apparently homothetic. Right: metric entropy. Linear dependences $h \approx \beta(T_E - T_E^{crit})$ occur for all values of JT. (A color version of this figure appears between pages 196 and 197).

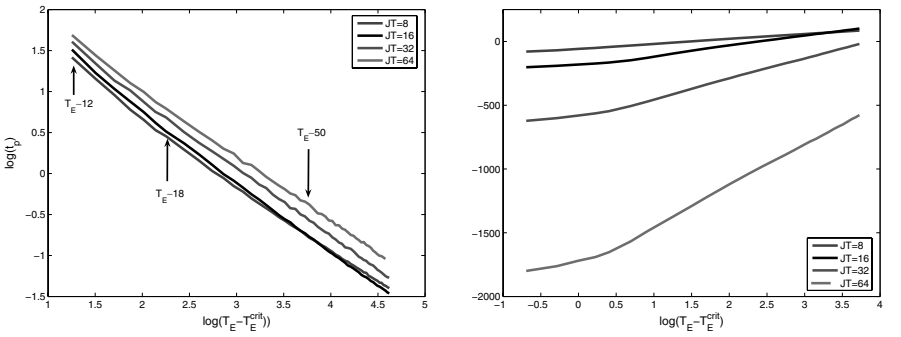


Fig. 12.2. Left: Log-Log plot of $t_p = 1/\lambda_1$ versus $T_E - T_E^{crit}$. Right: Volume of the Bounding Box V_{BB} versus $T_E - T_E^{crit}$. For a description of the power-law fits, see text. (A color version of this figure appears between pages 196 and 197).

for low-dimensional cases such as the Henon (Wang and Young 2001) or Lorenz (Viana 2000) strange attractors. We *assume* the existence of a *unique* SRB measure, coherently with the chaotic hypothesis proposed by Gallavotti and Cohen (1996) and Cohen and Gallavotti (1999), and refer to the sum of the positive Lyapunov exponents as metric entropy.

The metric entropy is presented in Fig. 12.1 right, whereas the mean predictability time for infinitesimal perturbations $t_p = \lambda_1^{-1}$ is plotted in Fig. 12.2 left. The metric entropy has a marked linear dependence $h \approx \beta(T_E - T_E^{crit})$, with β ranging from $\approx$ 0.15 (JT = 8) to $\approx$ 0.5 (JT = 64). Moreover, for a given value of T_E, the metric entropy increases with JT. Moreover, for fixed JT the predictability time decreases monotonically with T_E, whereas it increases for increasing values of JT. We note that, for all values of JT, if $T_E > 14$ we have that $t_p < 10$, which corresponds in physical units to $t_p < 12$ days. Moreover, in the range $T_E \geq 12$, t_p is proportional to $(T_E - T_E^{crit})^{\gamma}$, with γ within $[-0.85,-0.8]$ depending on the considered value of JT. From the dynamical viewpoint, this means on one hand that the maximal sensitivity of the system to variations in the initial condition along a *single direction* is largest for JT = 8. On the other hand, there are more *active degrees of freedom* for JT = 64 and they collectively produce a faster *forgetting* of the initial condition with time.

A more precise assessment of the time scales of the system and of predictability fluctuations may be gained by computing generalized Lyapunov exponents and performing the related multifractal diagnostics (Benzi et al. 1985, Paladin and Vulpiani 1987, Crisanti et al. 1993). which is outside the scope of this work.

3.2 Bounding box of the attractor

The bounding box of a set of points in an N-dimensional space is defined as the smallest hyperparallelepiped containing the considered set (Smith 2000). When the Hadley equilibrium is the attractor, the volume $V_{BB} = 0$, while it is non-zero if the computed orbit is attracted to a periodic orbit, a two-torus or a strange attractor. In all

cases V_{BB}, which measures the bulk size of the attractor in phase space, grows with T_E. Actually, each of the factors in the product increases with T_E, so that expansion occurs in all directions of the phase space. This matches the expectations on the behavior of a dissipative system having a larger input of energy.

In the right panel of Fig. 12.2 we present a plot of $\log(V_{BB})$ as function of T_E for the selected values of $JT = 8$, 16, 32, and 64. In the case $JT = 8$, V_{BB} obeys with great precision the power-law $V_{BB} \propto (T_E - T_E^{crit})^{\gamma}$ in the whole domain $T_E \geq 9$. The best estimate for the exponent is $\gamma \sim 40$. Given that the total number of Fourier components is $6 \times JT = 48$, this implies that the growth of the each side of the bounding box is on the average proportional to about the $5/6^{th}$ power of $(T_E - T_E^{crit})$. For $JT \geq 16$, two distinct and well defined power-law regimes occur:

$$V_{BB} \propto (T_E - T_E^{crit})^{\gamma}, \quad \gamma = \varepsilon N \quad \varepsilon \sim \begin{cases} 1/3, & T_E - T_E^{crit} \leq 1.5, \\ 5/6, & T_E - T_E^{crit} \geq 1.5, \end{cases} \tag{8}$$

where $N = 6 \times JT$ is the number degrees of freedom. We emphasize that in all cases the uncertainties on γ, which have been evaluated with a standard bootstrap technique, are rather low and total to less than 3% of the best estimate of γ. Moreover, the uncertainty of the power-law fit greatly worsens if we detune the value of T_E^{crit} by, say, 0.3, thus reinforcing the idea that it is meaningful to fit a power-law against the logarithm of $(T_E - T_E^{crit})$.

The proportionality between the exponent and the number of degrees of freedom suggests considering separately the various sides of the bounding box hyperparallelepiped, *i.e.*, each of the factors in the product. For all values of *JT*, each side increases as $(T_E - T_E^{crit})^{\gamma/N}$, so that the hyperparallelepiped obeys a sort of self-similar scaling with T_E.

The comparison for various values of *JT* of factors in corresponding to the same spectral component for the same value of $T_E - T_E^{crit}$ provides insight about the sensitivity to model resolution. The factors related to the the gravest latitudinal modes, such as agree with high precision, thus suggesting that the large scale behavior of the system is only slightly affected by variation of model resolution. When considering the terms related to the fastest latitudinally varying modes allowed by the lower resolution model, those obtained for *JT* = 32 are *larger* than the corresponding factors obtained for *JT* = 64. This is likely to be the effect of spectral aliasing: the dynamics contained in the scales which are resolved in the higher-resolution models are projected in the fastest modes of the model with lower resolution.

4 Statistics of the Total Energy and Zonal Wind

4.1 Total energy

The total energy of the system $E(t)$ is a global observable of obvious physical significance, statistically obeying a balance between the external forcing and the

internal and surface dissipation. The horizontal energy density of the two-layer QG system can be expressed as follows:

$$e(x,y,t)=\frac{\delta p}{g}\left[\left(\vec{\nabla}\phi\right)^2+\left(\vec{\nabla}\tau\right)^2+\frac{2}{H_2^2}\tau^2\right]. \tag{9}$$

Here the factor $\delta p/g$ is the mass per unit surface in each layer, the last term and the first two terms inside the brackets represent the potential and kinetic energy, respectively, thus featuring a clear similarity with the functional form of the energy of a harmonic oscillator. Note that in Eq. (10) the potential energy term is half of what reported in Pedlosky (1987), which contains a trivial mistake in the derivation of the energy density, as discussed with the author of the book. The total energy $E(t)$ is evaluated by integrating the energy density expression:

$$E(t)=\int_0^{L_y}dy\int_0^{L_x/6}dxe(x,y,t)=6\int_0^{L_y}dy\int_0^{L_x/6}dxe(x,y,t) \tag{10}$$

Potential energy is injected into the system by the baroclinic forcing τ^*. Part of it is transformed into wave kinetic energy by baroclinic conversion, and kinetic energy is eventually dissipated by friction – in our case, Ekman pumping. This constitutes the Lorenz energy cycle (Lorenz 1955), analized for this system in Speranza and Malguzzi (1988). In Tab. 1 we report the conversion factor for $E(t)$ between the non-dimensional and dimensional units. For the Hadley equilibrium, we have:

$$\overline{E(t)}=\frac{\delta p}{g}L_xL_y\left(\frac{RT_E}{4f_0}\frac{1}{1+\frac{\kappa}{\nu_E}\left(\frac{\pi}{L_y}\right)^2}\right)^2\left(\frac{\pi^2}{L_y^2}+\frac{1}{H_2^2}\right). \tag{11}$$

The total energy is proportional to T_E^2 and is mostly (≈95%) stored as potential energy, which is described by the second term of the sum in Eq. (11). In Fig. 12.3 left we present the results obtained for the various values of *JT* used in this work. In the left panel we present the *JT* = 64 case, which is representative of what obtained also in the other cases. The time-averaged total energy is monotonically increasing with T_E, but when the system enters the chaotic regime, $\overline{E(t)}$ is much lower than the value for the coexisting Hadley equilibrium. This behavior may be related to the much larger dissipation fueled by the activation of the smaller scales. In the chaotic regime $E(t)$ is characterized by temporal variability, which becomes more and more pronounced for larger values of T_E. The overall agreement among the results obtained for $\overline{E(t)}$ by choosing various *JT* values is good but progressively worsens when decreasing *JT*: for *JT* = 32, the maximal fractional difference is less than 0.01, while for *JT* = 8 it is about one order of magnitude larger. Differences among the representations given by the various truncations levels also emerge in power-law fits such as $\overline{E(t)}\propto T_E^{\gamma}$. In the regime where the Hadley equilibrium is attracting, this fit is exact, with exponent $\gamma=2$. For $T_E-T_E^{crit}\le 1.5$ and $T_E>T_E^H$ (above the first Hopf bifurcation), for all the values of *JT* the power-law fit is good, with

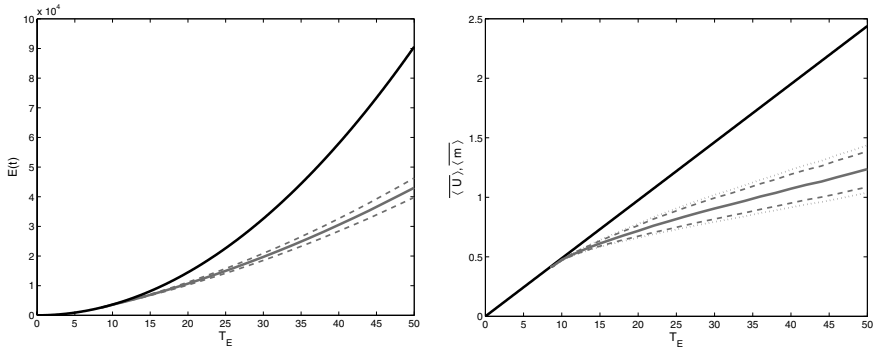


Fig. 12.3. Left: $\overline{E(t)}$ obtained from model integrations for JT = 64 (magenta line) and for Hadley equilibrium (black line). Right: $\overline{\langle U(t)\rangle} = \overline{\langle m(t)\rangle}$ obtained from model integrations for JT = 64 (magenta line) and for Hadley equilibrium (black line). See text for details. (A color version of this figure appears between pages 196 and 197).

$\gamma = 1.90 \pm 0.03$, so that a weakly subquadratic growth is realized. For $T_E - T_E^{crit} \geq 1.5$, only the *JT* = 32 and 64 simulations of $\overline{E(t)}$ obey with excellent approximation a weaker power-law, with $\gamma = 1.52 \pm 0.02$ in both cases, while the cases *JT* = 8 and 16 do not satisfactorily fit any power-law. The agreement worsens in the upper range of T_E, which points at the criticality of the truncation level when strong forcings are imposed. Nevertheless, the observed differences are strikingly small between the cases, say, *JT* = 8 and *JT* = 64, with respect to what could be guessed by looking at the previously analyzed diagnostics such as Lyapunov dimension, entropy production, and bounding box volume.

4.2 Zonal wind

We consider $U(y,t) = 1/2\big(u_1(y,t) + u_3(y,t)\big)$ and $m(y,t) = 1/2\big(u_1(y,t) - u_3(y,t)\big)$, which represent the zonal average of the mean and of the halved difference of the zonal wind at $p = p_1$ and $p = p_3$ at latitude *y*, respectively. The latitudinal average $\langle U(y,t)\rangle$ is then proportional to the total zonal momentum of the atmosphere, whereas $\langle m(y,t)\rangle$, by geostrophic balance, is proportional to the equator-to-pole temperature difference. Computation of the latitudinal average at the time-independent Hadley equilibrium is straightforward:

$$\big\langle m(y)\big\rangle = \big\langle U(y)\big\rangle = \frac{R}{f_0 L_y}\frac{T_E}{2}\frac{1}{1+\frac{\kappa}{\nu_N}\left(\frac{\pi}{L_y}\right)^2}. \tag{12}$$

Since we cannot have net, long-term zonal forces acting on the atmosphere at the surface interface, the spatial average of the zonal wind at the $p = p_3$ must be zero. Therefore, the outputs of the numerical integrations must satisfy the constraint $\big\langle m(y,t)\big\rangle = \big\langle U(y,t)\big\rangle$, where $\overline{X}$ denotes the time-average of the field *X*. In the right panel of Fig. 13.3 we plot the outputs for *JT* = 64, which, similarly to the total energy case, is well representative of all the *JT* cases. The constraint is obeyed within

numerical precision. The average winds are monotonically increasing with T_E, but, when the system enters the chaotic regimes, the averages $\langle m(y,t)\rangle = \langle U(y,t)\rangle$ have a smaller value than at the corresponding Hadley equilibrium, and they display sub-linear growth with T_E. Moreover, for $T_E > T_E^{crit}$ the temporal variability of the time series $\langle m(y,t)\rangle$ and $\langle U(y,t)\rangle$ increases with T_E. The variability of $\langle m(y,t)\rangle$ is larger than that of $\langle U(y,t)\rangle$, probably because the latter is related to a *bulk* property of the system such as the total zonal momentum.

The overall agreement between the various truncation levels progressively worsens for smaller *JT*, similarly to what observed for the total energy of the system. In particular, differences emerge as one attempts power-law fits of the form $\langle m(y,t)\rangle = \langle U(y,t)\rangle \propto T_E^{\gamma}$. For the Hadley equilibrium regime we have $\gamma = 1$. For $T_E \leq 10$ and above the first Hopf bifurcation, for all values of *JT* the power-law fit is good, with $\gamma = 0.875 \pm 0.005$. For $T_E - T_E^{crit} \geq 1.5$, only the simulations with $JT = 32$ and 64 obey a power-law (with $\gamma = 0.58 \pm 0.02$) with excellent approximation.

This implies that while the time-averaged meridional temperature difference between the northern and southern boundary of the system increases monotonically with T_E, as to be expected, the realized value is greatly reduced by the onset of the chaotic regime with respect to the corresponding Hadley equilibrium. This is the signature of the negative feedback due to the following mechanism: when the poleward eddy transport of heat is realized, it causes the reduction of the meridional temperature gradient, thus limiting by geostrophy the wind shear, which causes in turn a reduction of baroclinically-induced eddies *towards* the marginal stability. This process can be considered to be the statistical generalization of the classical *baroclinic adjustment* (Stone 1978). The latter implies that the system balances near an average state which corresponds to a fixed point which is neutral with respect to baroclinic instability. Theoretical justifications for this type of adjustment often rely on the idea that a variational principle holds for the relationship between the basic state gradient and the heat and momentum fluxes. This is the case, *e.g.*, of classical convection, for which it can be proved that the most unstable mode is the one carrying heat most efficiently in the direction of the basic state gradient. In fact, such a variational principle does not hold for ordinary baroclinic instability, as it can be proved that the most rapidly growing baroclinic mode is not the one with the largest heat flux. But in this model, as opposed to the general case, the variational assumption in question is essentially correct, since only one zonal wave is considered (Speranza and Malguzzi 1988). Nevertheless, the adjustment mechanism does not keep the system *close to marginal stability* since for $T_E > T_E^{crit}$ both the instantaneous and the time-averaged fields of the system are rather different from those realized for $T_E \approx T_E^H$, near the edge of the Hadley equilibrium. This denies the possibility that the time-mean circulation is maintained by eddies which can be parameterized in terms of the time-mean fields.

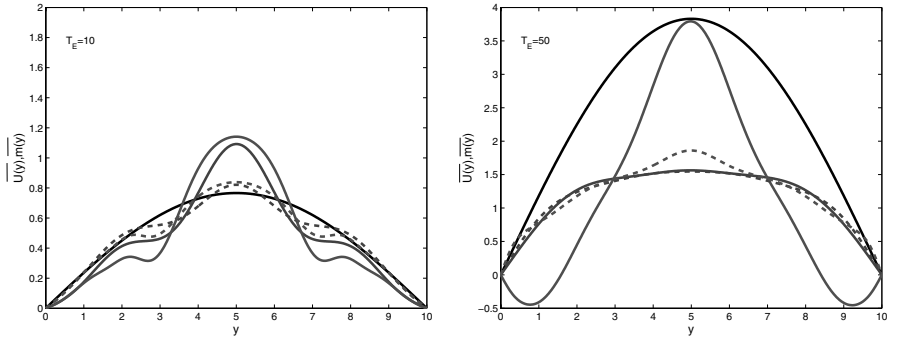


Fig. 12.4. Time-averaged profiles $\overline{U(y)}$ (solid lines) and $\overline{m(y)}$ (dashed lines). The black line indicates the $U(y) = m(y)$ Hadley equilibrium profiles. Red (blue) line refers to the $JT = 8$ ($JT = 32$) case. T_E is indicated. Vertical scale on the left figure is 1/2 of that of the right figure. (A color version of this figure appears between pages 196 and 197).

By examining more detailed diagnostics on the winds, such as the time-averaged latitudinal profiles $\overline{U(y)}$ and $\overline{m(y)}$ (Fig. 12.4), relevant differences are observed between $JT = 8$ and the other three cases. Results are presented for $JT = 8$ and $JT = 32$, the latter being representative also of $JT = 16$ and 64. We first note that already for $T_E = 10$, where only a weakly chaotic motion is realized, the $\overline{U(y)}$ and $\overline{m(y)}$ profiles feature in both resolutions relevant qualitative differences with respect to the corresponding Hadley equilibrium profile, although symmetry with respect to the center of the channel is obeyed. The $\overline{U(y)}$ and $\overline{m(y)}$ profiles are different (the constraint $\langle \overline{m(y,t)} \rangle = \langle \overline{U(y,t)} \rangle$ being still satisfied), with $\overline{U(y)} > \overline{m(y)}$ at the center and $\overline{U(y)} < \overline{m(y)}$ at the boundaries of the channel. Nevertheless, as for the Hadley equilibrium, both $\overline{U(y)}$ and $\overline{m(y)}$ are positive and are larger at the center of the channel than at the boundaries. Consequently, at $p = p_1$ there is a westerly flow at the center of the channel and easterly flows at the two boundaries. At $p = p_3$ the wind is everywhere westerly and peaks at the center of the channel. Such features are more pronounced for the $JT = 32$ case, where the mechanism of the convergence of zonal momentum is more accurately represented.

For larger values of T_E, the differences between the two truncation levels become more apparent. For $JT = 8$, the observed $\overline{U(y)}$ and $\overline{m(y)}$ profiles tend to flatten in the center of the channel and to become more similar to each other. Therefore, somewhat similarly to the Hadley equilibrium case, the winds at $p = p_1$ tend to vanish and all the dynamics is restricted to $p = p_3$. The $\overline{m(y)}$ profiles for $JT = 32$ are quite similar to those of $JT = 8$, even if they peak and reach higher values in the center of the channel and are somewhat smaller at the boundaries. So when a finer resolution is used, a stronger temperature gradient is realized in the channel center. The $\overline{U(y)}$ profiles obtained for $JT = 32$ are instead very different. They feature a strong, well-defined peak in the channel center and negative values near the boundaries. Therefore, the winds in the upper pressure level are strong westerlies, and peak in the center of the channel, while the winds in the lower pressure level feature a relatively strong westerly jet in the center of the channel and two

compensating easterly jets at the boundaries. The fact that for higher resolution the wind profiles are less smooth and have more evident jet-like features is related to the more efficient mechanism of barotropic stabilization, which, through zonal wind convergence, *keeps the jet together* (Kuo 1973).

5 Summary, Conclusions and Future Developments

We have described the construction and the dynamical behavior of an *intermediate complexity* model of the atmospheric system. We take as prototypal model for the GAC a QG two-layer model where the mid-latitude atmosphere is taken as being composed of two-layers and the β-plane approximation is considered.

A single zonal wave solution is assumed and by a spectral discretization in the latitudinal direction, the latter equation is reduced to a system of $N = 6 \times JT$ ordinary differential equations, where $JT + 1$ is the number of nodes of the (latitudinally speaking) fastest varying base function. We have considered the cases $JT = 8$, 16 , 32, and 64. Although obviously relevant ingredients of geometrical (horizontal convergence due to the Earth curvature, latitudinal boundary conditions at the margins of the mid-latitudes, circumpolar vortex, etc.) and dynamical (stabilization mechanisms such as the so-called *barotropic governor* (Nakamura 1993) nature of the real atmospheric circulation are still missing in this simplified theoretical representation, the model features some fundamental processes determining the general circulation of the Earth atmosphere. In particular, the acting processes are the baroclinic conversion, transforming available potential energy into waves; the non-linear stabilization of the zonal jet by eddy momentum convergence from non-symmetric disturbances; and the thermal diffusion and Ekman pumping-driven viscous dissipation.

When a larger pool of available energy is provided, the dynamics of the system is richer, since the baroclinic conversion process can transfer larger amounts of energy to the disturbances. Therefore, by increasing the parameter T_E, which parameterizes the baroclinic forcing, the behavior of the system is greatly altered. The attractor changes from a fixed point to a strange attractor via a finite number of bifurcations, starting with a Hopf bifurcation at $T_E = T_E^H$ determining the destabilization of the Hadley equilibrium, and a final two-torus breakdown at $T_E = T_E^{crit}$.

The strange attractor is studied by means of the Lyapunov exponents. The Lyapunov spectra obtained for $JT = 32$, 64 resemble what obtained in more complex QG models (Vannitsem 1997; Snyder and Hamill 2003). A striking feature of this dynamical system is the smooth dependence on T_E of all of its dynamical properties. The metric entropy, representing the overall total dynamical instability, increases linearly with T_E for $T_E > T_E^{crit}$ for all examined values of *JT*, and increases with *JT*. The Lyapunov dimension D_L increases with both T_E and *JT*. In particular, by increasing T_E, initially the dimension grows with a sublinear power-law $D_L \propto (T_E - T_E^{crit})^{\gamma}$, followed by a linear scaling regime, while for large T_E, D_L *saturates*. The fact monotonical increase of the attractor dimension with the total energy of the system - both monotonically increasing with T_E - suggests that the system has a positive temperature, in a statistical mechanical sense. For $JT \geq 16$,

each side of the bounding box of the attractor increases as $\propto (T_E - T_E^{crit})^{1/3}$ for $T_E - T_E^{crit} \leq 1.5$ and as $\propto (T_E - T_E^{crit})^{5/6}$ for larger values of T_E, while for $JT = 8$ only the latter regime is present. Therefore, the ratios of the ranges of the various degrees of freedom remain unchanged when varying T_E, yielding a self-similar scaling.

When the system is in the chaotic regime, the average total energy and zonal winds have lower values than those of the corresponding unstable Hadley equilibrium, because the occupation of the faster-varying latitudinal modes fuels viscous dissipation, which acts preferentially on the small scales. The total energy and the average wind field obey with excellent approximation a subquadratic and sublinear power-laws $\propto T_E^{\gamma}$, respectively, and agree *quantitatively* for all values of *JT*, essentially because these quantities are representative of global balances.

If general enough, the scaling properties could be of great help in setting up a theory for the overall statistical properties of the GAC and in guiding - on a heuristic basis - both data analysis and realistic simulations. A leading example for this would be the possibility of estimating the sensitivity of the output of the system with respect to changes in the parameters. Physical insight into the relevant mechanisms can be obtained by considering the main feedback mechanisms setting the average statistical properties of the system, *e.g.* when changing the forcing T_E from $T_E = T_E^0 > T_E^{crit}$ to $T_E = T_E^0 + \Delta T_E$. The change in T_E directly causes an increase in the wind shear through Newtonian forcing and indirectly an increase in the value of the mean zonal momentum through Ekman pumping. The increase in the wind shear enhances the baroclinic conversion and so the eddy components. In turn, the eddy stresses act as feedback by depleting baroclinity and by increasing the momentum through barotropic convergence, which closes the feedback loop. In the chaotic regime, these feedback mechanisms act as a statistical baroclinic adjustment process, which reduces the average value of the equator-to-pole temperature difference. This process is not reducible to the stationary balances of the classical theory of GAC, since it takes place when the system lives in a strange attractor. This situation certainly corresponds more closely, at least at the conceptual level, to the equilibration occurring in the atmosphere. Preliminary calculations performed by the authors together with Vannitsem (unpublished results) suggest that the power-laws presented here hold also on simplified yet global models of the atmospheric circulation, so the obtained scalings laws might be very helpful in establishing a sort of bulk *climate theory* for the atmospheric disturbances.

The fact that all the considered dynamical indicators and physical quantities feature a smooth, simple dependence with respect to the forcing parameter T_E and a limited variation is at first striking if one keeps in mind the phenomena typically occurring in low-dimensional systems. The latter often feature *attractor crises*, which are bifurcations in which the size and shape of a strange attractor drastically changes across a suitable parameter value (see Robert et al. (2000) and references therein). A particular type of crisis occurs when a saddle-node bifurcation takes place on the strange attractor, which is destroyed, and an attracting equilibrium or periodic orbit is created. This *phase-locked* attractor persists for an interval - *window of periodicity* - of parameter values, outside of which the strange attractor reappears.

Such windows of periodicity are known to occur in many systems and are even dense in the parameter space of certain one-dimensional models (Broer et al. 2002). Although bifurcations certainly occur as T_E varies in the fully chaotic regime (*e.g.*, at the values where a Lyapunov exponent crosses zero and becomes positive), these bifurcations do not induce sharp variations or crises. Moreover, we strongly doubt that there is a way to explain the smooth behaviour of the system by relating it to the bifurcation where the strange attractor is created, since the latter occurs for much lower values of T_E. Also compare the discussion in Speranza and Malguzzi (1988).

Moreover, when beyond a certain threshold of a parameter new physical processes come into play or a given process reaches a qualitatively distinct statistical balance, we do not expect to find a global uniform parametrically-controlled self-scaling of the system statistical properties. The change in the slope observed for value of $T_E - T_E^{crit} \approx 1.5$ may then be interpreted as a somewhat fundamental changeover from a quasi-linear baroclinic activity to a fully chaotic regime. A possible source of large - anyway likely to be smooth - and non-self-similar variations of the statistical properties of the attractor with respect to external parameters is resonant behavior, *i.e.* the preferential occurrence of certain physical processes for a bounded range of values of a given parameter. In the case of atmospheric dynamics, this effect might be more relevant than the occurrence of windows of periodicity, although in a range of variability which is different from what explored here, namely the low-frequency variability (Malguzzi and Speranza 1981; Benzi et al. 1986; Benzi and Speranza 1989; Ruti et al. 2006).

The chaotic hypothesis (Gallavotti and Cohen 1996; 1999) and the conjectures in (Albers and Sproot 2006) express similar concepts: despite the fact that the available theoretical and phenomenological evidence indicates that structural stability of systems having strange attractors is not typical, from the physical/numerical point of view typical chaotic systems behave *as though* they were structurally stable under perturbations if the dimension is sufficiently high. This fundamental assertion has both diagnostic and prognostic consequences: from the diagnostic viewpoint, the conjectures in (Albers and Sproot 2006) state that systems of the considered type display smooth-like dependence on external parameters; from the prognostic viewpoint, the chaotic hypothesis allows establishing fluctuation-dissipation theorems and Onsager relations in certain models for developed turbulence (Gallavotti and Cohen 1999). We propose that the windows of periodicity, if existing at all, are so narrow to be not detectable, and that the SRB measure varies continuously on the parameter set where a strange attractor exists.

An interesting approach is understanding the statistical properties of the waves as a function of the zonal wind fields, taken, as a first approximation, as external stochastic inputs with given statistical properties parameterized in terms of T_E. This amounts to considering the zonal wind as a sort of integrator of the wave disturbances, which is consistent with the almost-linearity of the evolution equations of the wave components of the model considered in this study. Such a framework, which assumes the possibility of neglecting the direct wave-wave interactions, seems to be well justified (Schneider 2006). The evolution equations for the wave fields,

indicated as $\vec{x}$, can be represented to be of the form $\dot{\vec{x}} = M(t)\vec{x}$, where M is a random square matrix (Crisanti et al. 1993) whose statistical properties depend on those of the U and m fields. Therefore, the study of the statistics of M, *e.g.* eigenvalues, eigenvectors, singular values, singular vectors, as a function of T_E may inform us on the properties of the generated waves, which can be associated to the destabilizing and stabilizing physical processes. The acknowledgement of the inadequacy of the paradigm of neutral atmospheric waves and of the need of a detailed description of the statistics of growth and decay of the waves would have far-reaching impacts. In fact, when considering the spectral representation of the mid-latitude waves of the real or *realistic* (*i.e.* resulting from models such as GCMs) atmosphere, we find 2D wavenumber-frequency spectral densities that only vaguely correspond, in a statistical sense, to well-defined dispersion relations $\omega = \omega(k)$, since at all frequencies the spectral width is relatively wide, essentially because of the finite life cycle of the waves (Dell'Aquila et al. 2005; Lucarini et al. 2006). Moreover, the proposed approach might serve also to understand at mathematical level the power-law scalings and thus making a sense of the self-similarity of the system's properties. Note that due to the absence of direct wave-wave interactions, this approach is somewhat dual with respect to a turbulence-oriented approach.

Further main, necessary directions of future work along the proposed line of approach to GAC can be summarized as follows:

- Analyzing the debated properties in more articulated numerical models. An important issue may be the separation of the properties of the baroclinic mid-latitude jet system from the low-frequency variability; using more realistic atmospheric models comprising bottom orography, a physically relevant experiment having the purpose of finding the break-up of self-similarity is to parameterically tune and detune the onset of the orographic baroclinic energy conversion (Buzzi et al. 1984, Ruti et al. 2006).
- Consolidating and explaining the knowledge - see Schneider (2006) – that in the real atmosphere the dynamics of traveling baroclinic disturbances is dominated by the *wave-zonal flow* interaction and the role of the *wave-wave* interaction is minor. This may be crucial for modeling in detail the extratropical storm tracks (Hoskins and Valdes 1990).

References

Albers, D. J. and Sprott, J. C. (2006) Structural stability and hyperbolicity violation in high-dimensional dynamical systems. Nonlinearity 19, 1801–1847.

Benzi, R., Paladin, G., Parisi, G. and Vulpiani, A. (1985) Characterisation of intermittency in chaotic systems. J. Phys. A. 18, 2157–2165.

Benzi, R., Malguzzi, P., Speranza, A. and Sutera, A. (1986) The statistical properties of general atmospheric circulation: observational evidence and a minimal theory of bimodality. Quart. J. Roy. Met. Soc. 112, 661–674.

Benzi, R. and Speranza, A. (1989) Statistical properties of low frequency variability in the Northern Hemisphere. J. Climate 2, 367–379.

Bonatto, C., Garreau, J. C. and Gallas, J.A.C. (2005) Self-similarities in the frequency-amplitude space of a loss-modulated CO_2 laser. Phys. Rev. Lett. 95, 143905.

Broer, H.W., Simó, C. and Vitolo, R. (2002) Bifurcations and strange attractors in the Lorenz-84 climate model with seasonal forcing. Nonlinearity 15, 1205–1267.

Buzzi, A., Trevisan, A. and Speranza, A.(1984) Instabilities of a baroclinic flow related to topographic forcing. J. Atmos. Sci. 41, 637–650.

Charney, J.G. (1947) The Dynamics of Long Waves in a Baroclinic Westerly Current. J. Atmos. Sci. 4, 136–162.

Charney, J.G. (1948) On the scale of atmospheric motions. Geofys. Publik. 17, 251–265.

Charney, J.G. and Devore, J.C. (1979) Multiple flow equilibria in the atmosphere and blocking. J. Atmos. Sci. 36, 1205–1216.

Cohen, E.G.D. and Gallavotti, G. (1999) Note on Two Theorems in Nonequilibrium Statistical Mechanics. J. Stat. Phys. 96, 1343–1349.

Crisanti, A. Paladin, G. and Vulpiani, A. (1993) Products of Random Matrices in Statistical Physics. Springer, New York.

Eady, E.T. (1949) Long waves and cyclone waves. Tellus 1, 33–52.

Eckmann, J. P. and Ruelle, D. (1985) Ergodic theory of chaos and strange attractors. Rev. Mod. Phys. 57, 617–655.

Dell'Aquila, A., Lucarini, V., Ruti, P.M. and Calmanti, S. (2005) Hayashi spectra of the northern hemisphere mid-latitude atmospheric variability in the NCEP–NCAR and ECMWF reanalyses. Clim. Dyn. 25, 639–652.

Dijkstra, H. A. (2005) *Nonlinear Physical Oceanography*. Springer, New York.

Farmer, J.D. (1982) Chaotic attractors of an infinite-dimensional dynamic system. Physica D 4, 366–393.

Farmer, J.D., Ott, E. and Yorke, J.A. (1983) The dimension of chaotic attractors. Physica D 7, 153–180.

Felici, M., Lucarini, V., Speranza, A., Vitolo, R. (2006) Extreme Value Statistics of the Total Energy in an Intermediate Complexity Model of the Mid-latitude Atmospheric Jet. Part I: Stationary case. J. Atmos. Sci., in press.

Foias, C. and Olson, E.J. (1996) Finite Fractal Dimension and Holder-Lipschitz Parametrization. Indiana Uni. Math. J. 45, 603–616.

Friz, P.K. and Robinson, J.C (2001) Parametrising the attractor of the two-dimensional Navier–Stokes equations with a finite number of nodal values. Physica D 148, 201–220.

Gallavotti, G. (1996) Chaotic hypothesis: Onsager reciprocity and fluctuation-dissipation theorem. J. Stat. Phys. 84, 899–926.

Ghil, M., Benzi, R. and Parisi, G. (Eds.) (1985) *Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*. North-Holland, Amsterdam.

Held, I.M. (2005) The Gap between Simulation and Understanding in Climate Modeling. Bull. Am. Meteor. Soc., 1609–1614.

Held, I.M. and Hou, A.Y. (1980) Nonlinear Axially Symmetric Circulations in a Nearly Inviscid Atmosphere. J. Atmos. Sci. 37, 515–533.

Hénon, M. and Pomeau, Y. (1976) Two strange attractors with a simple structure. In: Turbulence and Navier-Stokes equations 565. Springer-Verlag, 29–68.

Hoskins, B.J. and Valdes, P.J. (1990) On the existence of storm-tracks. J. Atmos. Sci. 47, 1854–1864.

Kalnay, E. (2003) *Atmospheric modeling, data assimilation and predictability*. Cambridge Univ. Press, Cambridge.

Kaplan, J. and Yorke, J. (1979) Chaotic behaviour of multidimensional difference equations. In: H.-O. Peitgen and H.-O. Walther (Eds.), *Functional Differential Equations and Approximations of Fixed Points*. Springer, Berlin, pp. 204–227.

Kubo, R. (1966) The Fluctuation Dissipation Theorem. Rep. Prog. Phys. 29, 255–284.

Kuo, H.L. (1973) Dynamics of quasigeostrophic flows and instability theory. Adv. Appl. Mech. 13, 247–330.

Leith, C.E. (1975) Climate response and fluctuation dissipation. J. Atmos. Sci. 32, 2022-2026.

Lorenz, E.N. (1955) Available potential energy and the maintenance of the general circulation. Tellus 7, 157–167.

Lorenz, E.N. (1967) *The Nature and Theory of the General Circulation of the Atmosphere*. World Meteorol. Organ., Geneva.

Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. Tellus 21, 289–307.

Lorenz, E.N. (1979) Forced and Free Variations of Weather and Climate. J. Atmos. Sci. 36, 1367–1376.

Lorenz, E.N. (1980) Attractor sets and quasi-geostrophic equilibrium. J. Atmos. Sci. 37, 1685–1699.

Lucarini, V. (2002) Towards a definition of climate science. Int. J. Environ. Pollut. 18, 409–414.

Lucarini, V., Saarinen, J.J., Peiponen, K.-E. and Vartiainen, E. (2005) *Kramers-Kronig Relations in Optical Materials Research*. Springer, Heidelberg.

Lucarini, V., Calmanti, S., dell'Aquila, A., Ruti, P.M. and Speranza, A. (2006) Intercomparison of the northern hemisphere winter mid-latitude atmospheric variability of the IPCC models. Clim. Dyn. DOI: 10.1007/s00382-006-0213-x.

Malguzzi, P. and Speranza, A. (1981) Local Multiple Equilibria and Regional Atmospheric Blocking. J. Atmos. Sci. 9, 1939–1948.

Malguzzi, P., Trevisan, A. and Speranza, A. (1990) Statistics and predictability for an intermediate dimensionality model of the baroclinic jet. Ann. Geoph. 8, 29–35.

Mantovani, R. and Speranza, A. (2002) Baroclinic instability of a symmetric, rotating, stratified flow: a study of the nonlinear stabilisation mechanisms in the presence of viscosity. Nonlinear Processes in Geophysics 9, 487–496.

Margules, M. (1903) Die energie der Stürme. Jahrb. Zentralanst. Meteor. Wien 40, 1–26.

Nakamura, N. (1993) Momentum flux, flow symmetry, and the nonlinear barotropic governor. J. Atmos. Sci. 50, 2159–2179.

Oseledec, V.I. (1968) A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems. Trudy Mosk. Mat. Obsc. (Moscow Math. Soc.) 19, 197pp.

Paladin, G. and Vulpiani, A. (1987) Anomalous scaling laws in multifractal object. Phys. Rep. 156, 147–225.

Pedlosky, J. (1987) *Geophysical Fluid Dynamics*. Springer, New York

Phillips, N.A. (1954) Energy transformations and meridional circulations associated with simple baroclinic waves in a two-level, quasi-geostrophic model. Tellus 6, 273–286.

Randriamampianina, A., Früh, W.-G., Maubert, P. and Read, P.L. (2005) DNS of bifurcations to low-dimensional chaos in an air-filled rotating baroclinic annulus. Preprint at http://www-atm.physics.ox.ac.uk/user/read/.

Robert, C., Alligood, K. T., Ott, E. and Yorke, J. A. (2000) Explosions of chaotic sets. Physica D, 144(1-2), 44–61.

Ruelle, D. (1990) Deterministic chaos: the science and the fiction. Proc. R. Soc. London A 427, 241–248.

Ruelle, D. (1998) General linear response formula in statistical mechanics, and the fluctuation-dissipation theorem far from equilibrium. Phys. Letters A 245, 220-224.

Ruti, P.M., Lucarini, V., Dell'Aquila, A., Calmanti, S. and Speranza, A. (2006) Does the subtropical jet catalyze the mid-latitude atmospheric regimes? Geophys. Res. Lett. 33, L06814.

Schertzer, D. and Lovejoy, S. (2004) Uncertainty and Predictability in Geophysics: Chaos and Multifractal Insights. In: R. S. J. Sparks and C. J. Hawkesworth (Eds.)*, The State of the Planet: Frontiers and Challenges in Geophysics*. American Geophysical Union, Washington. pp. 317-334

Schneider, V. (2006) The general circulation of the atmosphere. Ann. Rev. Earth Plan. Sci 34, 655–688, DOI:10.1146/annurev.earth.34.031405.125144.

Smith, L.A. (2000) Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems. In: A. Mees (Ed.), *Nonlinear Dynamics and Statistics*, Birkhauser, Boston, pp. 31–64.

Snyder, C. and Hamill, T.M. (2003) Leading Lyapunov vectors of a turbulent baroclinic jet in a quasigeostrophic model. J. Atmos. Sci. 60, 683–688.

Speranza, A. and Lucarini, V. (2005) Environmental Science: physical principles and applications. In: Encyclopedia of Condensed Matter Physics, F. Bassani, J. Liedl, P. Wyder eds., Elsevier, Amsterdam, in press.

Speranza, A. and Malguzzi, P. (1988) The statistical properties of a zonal jet in a baroclinic atmosphere: a semilinear approach. Part I: two-layer model atmosphere. J. Atmos. Sci. 48. 3046–3061.

Stone, P.H. (1978) Baroclinic adjustment. J. Atmos. Sci. 35, 561–571.

Vannitsem, S. and Nicolis, C. (1997) Lyapunov vectors and error growth patterns in a T21L3 quasigeostrophic model. J. Atmos. Sci. 54, 347–361.

Viana, M. (2000) What's new on Lorenz strange attractors? Math. Intelligencer 22-3, 6–19.

Wang, Q. and Young, L.-S. (2001) Strange Attractors with One Direction of Instability. Comm. Math. Phys. 218, 1–97.

13 Hindcast AGCM Experiments on the Predictability of Stratospheric Sudden Warming

Hitoshi Mukougawa[1], Toshihiko Hirooka[2], Tomoko Ichimaru[3] and Yuhji Kuroda[4]

[1] Kyoto University, Disaster Prevention Research Institute, mukou@dpac.dpri.kyoto-u.ac.jp
[2] Kyushu University, Department of Earth and Planetary Sciences, hirook@geo.kyushu-u.ac.jp
[3] Kyushu University, Department of Earth and Planetary Sciences, ichimaru@geo.kyushu-u.ac.jp
[4] Meteorological Research Institute, kuroda@mri-jma.go.jp

Abstract. High sensitivity to the initial condition of the prediction for a stratospheric sudden warming (SSW) event occurring in December 2001 reported by Mukougawa, Sakai, and Hirooka (2005) is confirmed by conducting a series of hindcast experiments using an atmospheric general circulation model (MRI/JMA-GCM). Similar precursory circulation anomaly in the troposphere for the occurrence of the SSW is also obtained through a regression analysis on the GCM experiments. Furthermore, it is revealed that the response of stratospheric circulation to the magnitude of the precursory anomaly is nonlinear, which suggests the existence of a threshold magnitude of the precursory anomaly for the occurrence of the SSW. Detailed investigation on the precursory event will enable us to reveal the dynamical relationship between tropospheric circulation anomaly and the subsequent SSW.

1 Introduction

Predictability studies on large-scale atmospheric motions have been mainly dealing with the tropospheric circulations (e.g., Kimoto, Mukougawa, and Yoden 1992), and interests on the predictability of the stratospheric motions are quite limited except for some pioneering works to simulate stratospheric sudden warming (SSW) events (e.g. Miyakoda, Strickler, and Hembree 1970; Mechoso, Yamazaki, Kitoh, and Arakawa 1985; Mechoso, Suarez, Yamazaki, Kitoh, and Arakawa 1986). SSWs are the most spectacular planetary-scale event in the wintertime stratospheric circulation. However, the predictability of the SSW has recently attracted much attention by upsurging interests on the downward influence of the stratospheric circulation change into the troposphere (Christiansen 2003; Reichler, Kushner, and Polvani 2005).

Mukougawa and Hirooka (2004) first reported prolonged predictability of a SSW event occurring in Dec 1998 up to 1 month by investigating 1-month forecasts issued once a week by the Japan Meteorological Agency (JMA). However, since their results were based upon control forecasts starting from initial conditions without any perturbation, the practical predictability of the SSW could not be assessed precisely. Then, Mukougawa, Sakai, and Hirooka (2005) (hereafter referred to as M05) examined

practical predictability of a SSW event occurring in Dec 2001 using all ensemble members of the JMA 1-month forecasts. They found that the SSW is predictable from at least 2 weeks in advance. Moreover, they revealed that the predictability of the SSW depends on the initial time of the forecast, and found the existence of high sensitivity to the initial condition for the SSW prediction during the onset period of the SSW. Similar variations of the predictability were also documented by Kimoto et al. (1992) for the forecast of a tropospheric blocking event.

The basic mechanism of SSW events has been already established by Matsuno (1971) in the framework of dynamical interactions between stratospheric zonal flows and upward propagating planetary waves from the troposphere. However, tropospheric precursors which generate and promote the upward propagation of planetary waves for the onset of the SSW have not been fully documented. The study of M05 detected a possible tropospheric precursor for the SSW event in Dec 2001 using a regression analysis for the JMA 1-month ensemble prediction. By this statistical analysis, they found that a characteristic zonal wind anomaly pattern in the upper troposphere associated with a persistent blocking event is significantly related to the subsequent occurrence of the SSW. However, since their work is based on a statistical analysis, we cannot confirm whether the statistically obtained precursor actually induces the following SSW.

Thus, in order to confirm the dynamical relevance of the obtained tropospheric precursor to the occurrence of the SSW event occurring in Dec 2001, we will conduct a series of hindcast experiments using an atmospheric general circulation model (MRI/JMA-GCM) in this study. The existence of high sensitivity to the initial condition for the SSW prediction reported by M05 will be also re-examined by the MRI/JMA-GCM experiments. We believe that detailed investigation on the precursory event by these hindcast experiments will enable us to reveal the dynamical relationship between tropospheric blocking event and the subsequent SSW.

2 Model

The model used in this study is a general circulation model (MRI/JMA-GCM) developed by the Japan Meteorological Agency (JMA) and Meteorological Research Institute (MRI) (Mizuta, Oouchi, Yoshimura, Noda, Katayama, Yukimoto, Hosaka, Kusunoki, Kawai, and Nakagawa 2006). The model is based on a global weather prediction model of the JMA (JMA-GSM0103), which was used to perform the JMA 1-month ensemble prediction during 2001/2002 winter. The 1-month forecast dataset was investigated in M05.

The computations were performed at a triangular truncation 96 with the linear Gaussian grid (TL96) in the horizontal. The model has 40 levels in the vertical, with the model top at 0.4 hPa. A semi-Lagrangian scheme is used to calculate the horizontal advection terms. The cumulus convection scheme proposed by Arakawa and Schubert (1974), and the level 2 turbulence closure scheme by Mellor and Yamada (1974) are implemented. Several physical processes important for the stratosphere, such as radiation, gravity wave drag, and a direct aerosol effect for the radiation processes are also implemented. The concentration of ozone is specified by the zonal-mean climatologi-

cal value during the computation. As boundary conditions, we used the monthly mean climatological sea surface temperature (SST) added with a constant SST anomaly from the climatology at the initial time. For further model details, the reader should consult JMA (2002) and Mizuta et al. (2006).

The initial conditions for the computations are identical to those used in M05, and are specified by the analysis added with initial perturbations in the JMA operational 1-month ensemble forecast. The JMA 1-month ensemble forecasts were conducted every Wednesday and Thursday from unperturbed initial condition and 12 perturbed conditions during 2001/2002 winter. The initial perturbations were obtained using the Breeding of Growing Modes (BGM) method (Toth and Kalnay 1993). They were applied to all pressure levels north of 20N with amplitude set to be 14.5 % of the climatological root-mean-square variance of 500-hPa heights. To verify model computations, JMA Global Analyses data set with 1.25-degree horizontal resolution at 23 levels from 1000 to 0.4 hPa is used.

3 Results

3.1 High sensitivity to the initial condition

At first, we checked the reproduction of high sensitivity to the initial condition observed prior to the onset of SSW shown in M05. The thick solid lines in Fig. 13.1 show the daily time series of the observed 10-hPa zonal mean temperature (T) at 80N. The temperature attains its peak on 28 Dec 2001. This warming is caused by the amplification of zonal wavenumber (WN) 1 planetary waves as seen in Fig. 13.2a which shows 3-day mean 10-hPa geopotential height during 27-29 Dec. The thin lines in Fig. 13.1 show the results of 30-day time integrations of MRI/JMA-GCM from initial conditions of the JMA ensemble 1-month forecasts starting from 5 and 6 Dec (Fig. 13.1a) and from 12 and 13 Dec (Fig. 13.1b). Some members successfully reproduce the occurrence of the SSW event for the computations from 5 and 6 Dec, while all the members starting from 12 and 13 well predict the warming episode as shown by M05. We also see that MRI/JMA-GCM tends to predict warmer temperatures in the polar stratospheric region for the SSW compared with the observation and the operational JMA 1-month forecasts (Fig. 13.1b). On the other hand, the forecasts starting from 5 and 6 Dec (Fig. 13.1a) have much larger spread among members than those from 12 and 13 Dec (Fig. 13.1b) during the mature phase of the SSW, in common with M05. Thus, we can also confirm high sensitivity to the initial condition for the prediction of the SSW during the onset period of the SSW by MRI/JMA-GCM hindcast experiments.

From Fig. 13.2, we can easily recognize a large spread among the ensemble members starting from 5 and 6 Dec. Figure 13.2b shows 3-day mean 10-hPa height field during 27-29 Dec for a forecast (run S) which attains the highest 10-hPa T at 80N on 28 Dec among the members, while Fig. 13.2c indicates a forecast (run F) with the lowest predicted temperature. For run S, the amplification of WN 1 component as for the analysis (Fig. 13.2a) is evident except for slight westward phase shift. For run F, however, the strong polar vortex still exists associated with low temperature in the

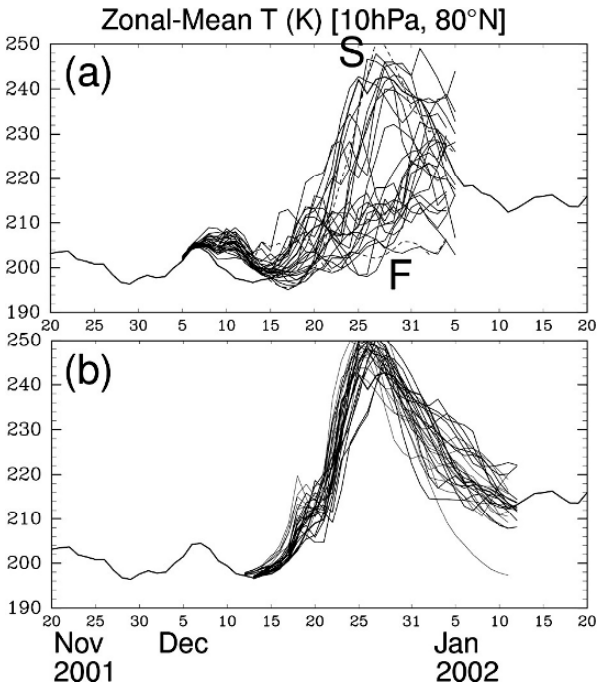


Fig. 13.1. Time variation of 10-hPa *T* at 80N from 20 Nov 2001 through 20 Jan 2002 for the analysis (thick solid lines), and for MRI/JMA GCM hindcast (thin solid lines) starting from 5 and 6 Dec 2001 (a), and 12 and 13 Dec 2001 (b). The dotted lines in (a) denote run S and run F.

polar region. Thus, the successful prediction for the amplification of WN 1 component is a key for the forecast of the SSW. As described in M05, the difference in the vertical component of WN 1 Eliassen-Palm (E-P) flux, which represents the vertical propagation of WN 1 wave activity (Andrews, Holton, and Leovy 1987), in higher latitudes of the lower stratosphere among ensemble members starting from 5 and 6 Dec becomes evident around 13 Dec. Therefore, we think that the period around Dec 13 corresponds to the onset of the SSW.

In order to find a necessary precursor for the SSW, we first compare the behavior of run S and run F during the onset period. Figure 13.3 shows 3-day mean 500-hPa geopotential height distributions for the onset period. The developed blocking high over the northeastern Atlantic shifts the Atlantic westerly jet considerably poleward in run S (Fig. 13.3b) as well as in the analysis (Fig. 13.3a). However, the blocking in run F (Fig. 13.3c) is much weaker, and the Atlantic jet resides near its climatological position around 60N. Since the blocking was already in a developing stage at the initial time of the hindcast experiment, weak persistence of the blocking is responsible for this difference. These features are also in common with those observed in M05.

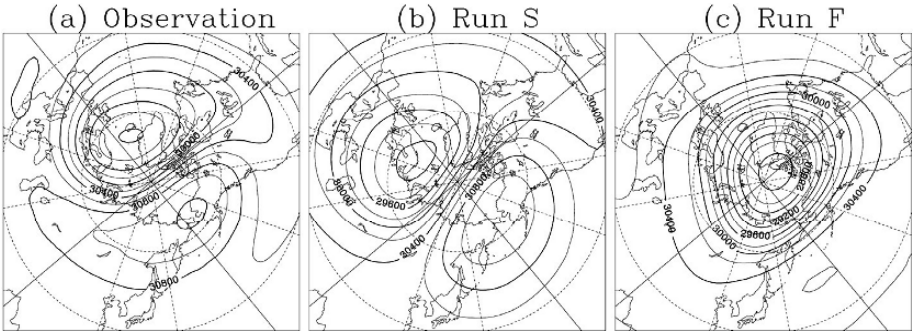


Fig. 13.2. 3-day mean 10-hPa height field (m) during 27-29 Dec 2001 for the analysis (a), run S (b), and run F (c). Contour interval is 200 m.

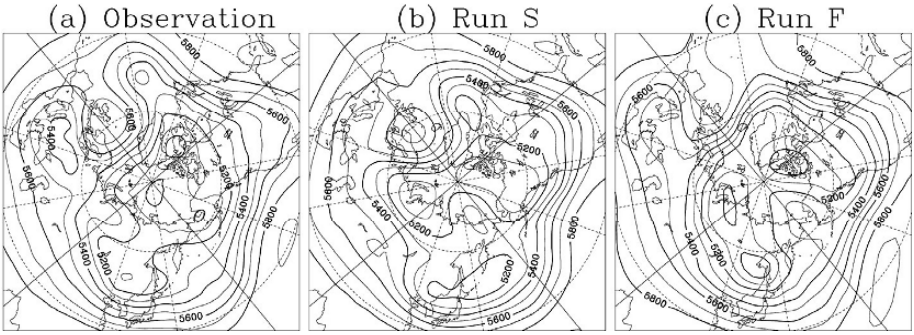


Fig. 13.3. 3-day mean 500-hPa height field (m) during 12-14 Dec 2001 for the analysis (a), run S (b), and run F (c). Contour interval is 100 m.

The meridional cross section of 3-day mean zonal-mean zonal wind (U) is shown in Fig. 13.4 for the analysis (Fig. 13.4a), run S (Fig. 13.4b), and run F (Fig. 13.4c) during the onset period. Since the configuration of U may affect the propagation and generation of planetary waves, it would be another important ingredient for the successful prediction of the SSW. Striking difference of run F from the analysis and run S is seen in high latitudes of the troposphere. For run F, weak easterlies exist around 80N with relatively strong westerlies around 60N in the upper troposphere. By contrast, a rather strong westerly jet in the upper troposphere is situated around 80N with weak westerlies to the south for the analysis and run S. The difference of U profile in high latitudes is associated with the persistence of the blocking high in the Atlantic sector as seen in Fig. 13.3. These features were also observed in the operational 1-month ensemble forecasts of JMA as shown by M05. It is also interesting to note that upper stratospheric U for the analysis and run S is much stronger than that for run F. Moreover, the axis of westerlies in the stratosphere for the analysis and run S shifts poleward in comparison with run F. The direction of WN 1 propagation and its wave activity are also

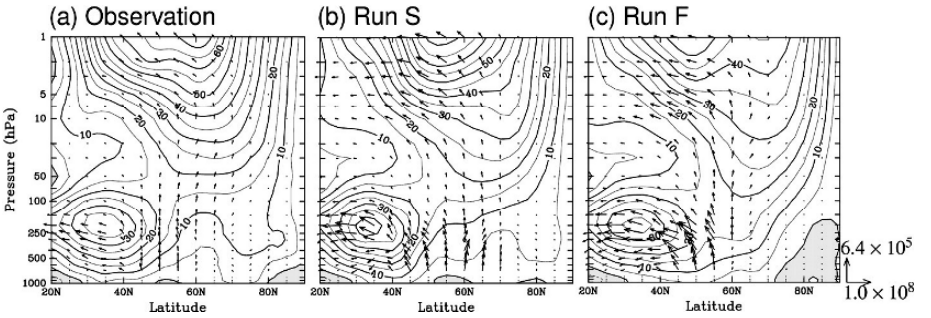


Fig. 13.4. Latitude-height cross sections of U (m/s) averaged over 12-14 Dec for the analysis (a), run S (b), and run F (c). The vectors show 3-day mean WN1 E-P flux (kg/s^2) above 700 hPa. E-P flux is scaled by the reciprocal square root of the pressure. The magnitude of the reference vectors at 1000 hPa is shown in the lower right corner.

shown by E-P flux vectors in Fig. 13.4. Although run S and run F represent larger wave activity of WN 1 compared with the analysis, the poleward and upward propagation of WN 1 component in run S is much enhanced in the troposphere around 60N compared with run F. This might suggest the importance of WN 1 generation in the troposphere for the SSW prediction.

3.2 Regression analysis

The possible relationship between the blocking high over the northeastern Atlantic and the subsequent occurrence of the SSW as seen in Fig. 13.3 is also confirmed by the regression analysis of 500-hPa height field using all ensemble (26) members of the hindcast experiments starting from 5 and 6 Dec as in M05. Figure 13.5 shows 500-hPa height anomalies averaged over 12-14 Dec associated with a one-standard deviation anomaly of 10-hPa T on 28 Dec, corresponding to the warming peak. Here the anomaly is defined as the deviation from the ensemble mean of the hindcast experiments (Fig. 13.5b). Statistical significance of correlations is also assessed by Student's *t-test* based on 24 degrees of freedom, and shown by shades. The degree of freedom is deduced by considering that 26 ensemble members are independent each other. From this figure, we confirm that the positive height anomaly over the northeastern Atlantic associated with the blocking is well correlated with the subsequent occurrence of the SSW after 2 weeks, consistent with the previous analysis on run S and run F. The correlated height anomaly center over the Atlantic in Fig. 13.5 possess almost the same horizontal position as that in Fig. 3a in M05 while other significant anomalies seen in Fig. 13.5a have no corresponding anomalies in M05, which again suggests an important role of the blocking for the SSW.

The similar regression analysis for the computed 3-day mean U during 12-14 Dec against 10-hPa T on 28 Dec is shown in Fig. 13.6. From this figure, we also confirm that the upper-tropospheric U anomaly around 60N associated with the positive height anomaly is significantly correlated with the occurrence of the SSW, which is also observed in M05. However, U anomaly in subtropical regions (polar regions) of Fig. 13.6a is significant (insignificant), in contrast with Fig. 4a in M05. Figure 13.6a also

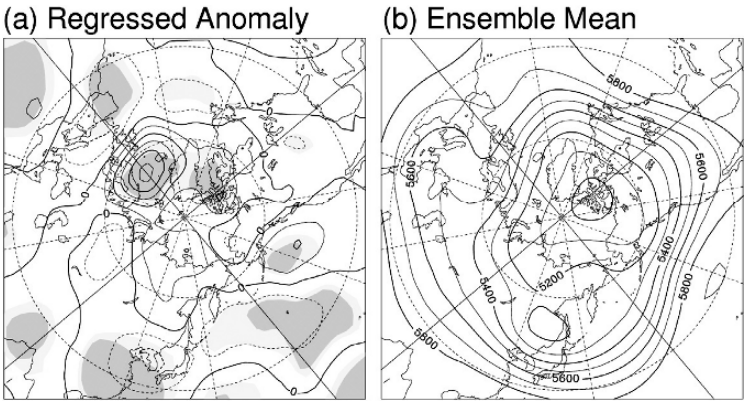


Fig. 13.5. (a) Regressed anomaly of the predicted 3-day mean 500-hPa height (m) during 12-14 Dec upon the predicted 10-hPa *T* at 80N on Dec 28 using all ensemble forecasts starting from 5 and 6 Dec by MRI-JMA GCM. The light (heavy) shades indicate regions where the statistical significance of the anomaly exceeds 95 (99)%. Contour interval is 20 m. (b) Ensemble average of the predicted 3-day mean 500-hPa height (m).

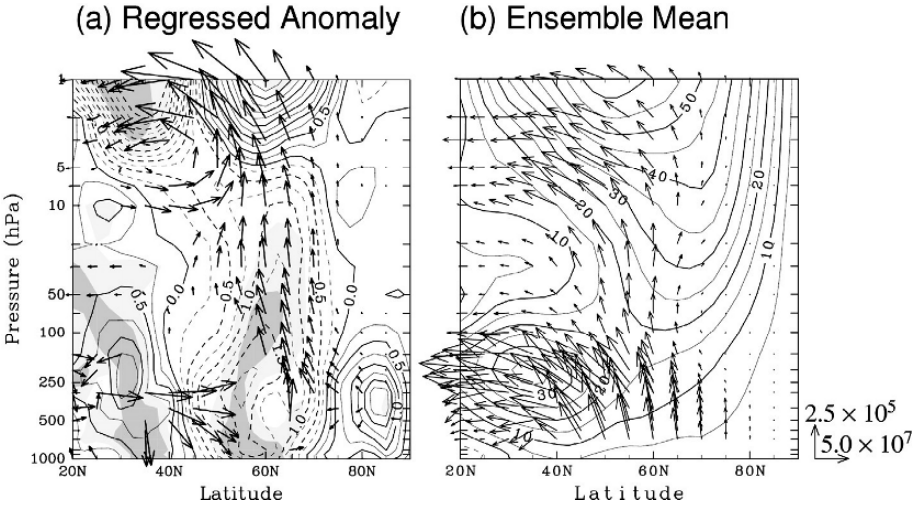


Fig. 13.6. As in Fig. 13.5, but for the predicted 3-day mean *U* (m/s) and WN 1 E-P flux (kg/s^2) during 12-14 Dec. The vectors in (a) shows the regressed WN 1 E-P flux anomalies of which vertical or meridional component is significant at 90% level above 700 hPa. Their magnitude is multiplied by 10. E-P flux is scaled by the reciprocal square root of the pressure. The magnitude of the reference vectors at 1000 hPa is shown in the lower right corner.

reveals that the poleward shift of upper stratospheric westerlies is also significantly correlated with the occurrence of the SSW as inferred from Fig. 13.4.

Figure 13.6a also shows the regressed anomaly of WN 1 E-P flux with statistical significance. Enhanced poleward propagation in the troposphere and active upward propagation from upper troposphere are intimately correlated with the SSW. The relationship is also in common with M05, but the statistical significance is much stronger compared with M05. Although the upward propagation of the ensemble averaged WN1 activity is most active around 60N in the lower stratosphere (Fig. 13.6b), the upward propagation is significantly enhanced around 70N. We also remark that the enhanced tropospheric generation of WN 1 component around 70N seems to be directly connected with the enhanced upward propagation to the stratosphere, which is not clearly seen in M05. These results would suggest that the generation rather than the propagation of WN 1 activity during the onset period is important for the occurrence of the SSW.

The high sensitivity to the initial condition for the SSW forecast found in Fig. 13.1a could be also inferred from Fig. 13.7 as in M05. This figure shows direction of the maximum spread among all ensemble members, represented by the regressed U anomalies on the leading principal component (PC1) of the computed 3-day mean U for 12-14 Dec of the hindcast experiments starting from 5 and 6 Dec. The EOF analysis was made for a domain from 1000 to 0.4 hPa poleward of 20N, and the data were weighted by the square root of the cosine of latitude as well as the square root of density at each level. The first EOF mode explains 43% of the total variance of U around the ensemble mean. The regressed U anomalies to PC1 have maxima in the upper troposphere, and are characterized by a barotropic tripole structure with nodes around 45N and 70N extending up to 10 hPa. In the upper stratosphere, a dipole structure of U anomaly with a node around 55N prevails. These features are very similar to the U profile correlated with the occurrence of the SSW shown in Fig. 13.6a. Thus, the coincidence could explain the reason why the high sensitivity to the initial condition is observed during the onset of the SSW as in M05. The regressed WN 1 E-P flux anomaly to PC1 as depicted by arrows in Fig. 13.7 also supports this conjecture. The upward propagation from the troposphere into the stratosphere around 60N is significantly enhanced as in Fig. 13.6a, which would trigger the SSW.

3.3 Hindcast Experiments using Regression Pattern

In order to reconfirm the important role of the statistically obtained regression pattern for the occurrence of the SSW shown in Figs. 13.5a and 13.6a, we performed a series of hindcast experiments starting from 13 Dec with several initial conditions composed of the ensemble mean field (Figs. 13.5b and 13.6b) and the regressed field multiplied by a coefficient α. The ensemble as well as the regressed field was computed for all prognostic variables of MRI/JMA-GCM, using predicted 3-day mean values for 12-14 Dec based on the GCM ensemble experiments starting from 5 and 6 Dec.

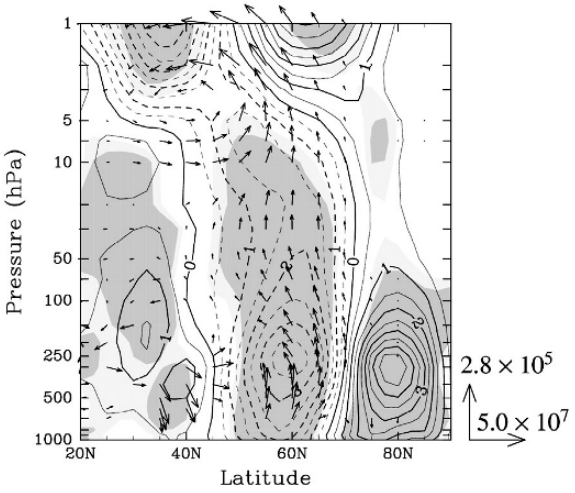


Fig. 13.7. As in Fig. 13.6a, but for the regressed *U* and E-P flux vectors upon PC1 of the predicted 3-day mean *U* during 12-14 Dec of the ensemble forecasts starting from 5 and 6 Dec.

The hindcast experiments were conducted for a range of the coefficient α from –2.0 to 4.0 with an increment of 0.5. The computation from the ensemble mean is given by $\alpha = 0$. We could expect from the regression analysis that positive α will enhance the possibility for the occurrence of a SSW while negative α will suppress SSW. The result of some hindcast experiments is shown by Fig. 13.8, which describes time evolution of the predicted *T* at 10 hPa and 80N. The numbers denoted on lines are α. From this figure, we find that positively large α actually tends to cause the warming while negatively large α tends to suppress the warming around 28 Dec. Thus, the important role of the regressed pattern shown by Figs. 13.5a and 13.6a for the occurrence of the SSW is reconfirmed by the hindcast experiments using MRI/JMA-GCM.

The relationship between α and predicted *T* at 10 hPa and 80N on 28 Dec is further examined by Fig. 13.9. The abscissa denotes α and the ordinate *T* at 10 hPa and 80N on 28 Dec. Crosses are for the hindcast experiments starting from 13 Dec while open circles are for MRI/JMA-GCM ensemble experiments starting from 5 and 6 Dec, and the solid circle denotes the observation. The value of *x*-axis for each experiment and the observation is given by the projection coefficient of *U(y,z)* anomaly field ***a*** onto the regressed field ***b*** defined by $\langle \boldsymbol{a},\boldsymbol{b}\rangle / \langle \boldsymbol{b},\boldsymbol{b}\rangle$, where $\langle \boldsymbol{a},\boldsymbol{b}\rangle$ denotes the inner product between vectors ***a*** and ***b*** over the region of 20N-90N and 1000-0.4 hPa taking account of the areal factor proportional to cosine of latitude. The anomaly is defined by the difference of 3-day mean *U* during 12-14 Dec from the ensemble averaged field shown by Fig. 13.6b. It is noticeable from this figure that there is overall tendency of

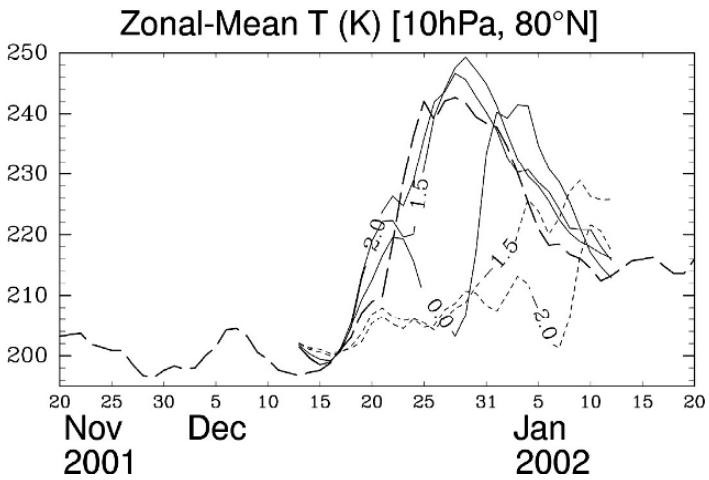


Fig. 13.8. As in Fig. 13.1, but for the MRI/JMA GCM hindcast experiments starting from 13 Dec 2001 with initial conditions composed of the ensemble mean field (Fig. 13.6b) and the regression field (Fig. 13.6a) multiplied by a coefficient α which is denoted on the lines. The analysis is shown by the broken line, predictions with positive (negative) coefficients by thin solid (dotted) lines.

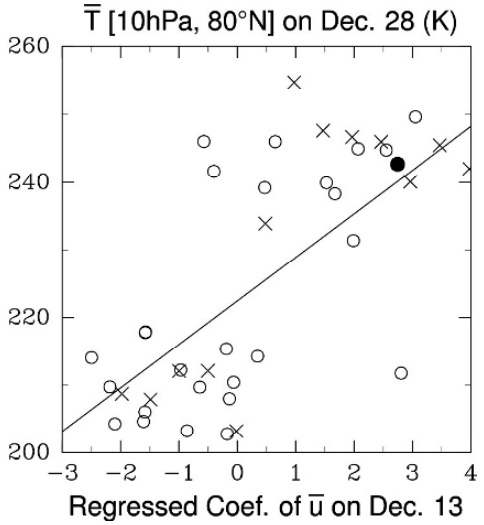


Fig. 13.9. Relationship between 10-hPa T (K) at 80N on Dec 28 and the coefficient α of 3-day mean U anomaly during 12-14 Dec. Open circles are for ensemble experiments starting from 5 and 6 Dec, crosses are for hindcast experiments from 13 Dec starting from the ensemble mean field (Fig. 13.6b) added with the regression field (Fig. 13.6a) multiplied by the coefficient α (value of the abscissa). The solid circle denotes the observation.

the increase of T on Dec. 28 with the increase of α, which is shown by a regression line (solid line) in Fig. 13.9. However, the relationship is well described by a stepwise function of α rather than a linear one. The computed T is clustered around 240 K (210 K) for positive (negative) values of α. This also suggests the existence of a threshold value of α for the occurrence of the SSW. Therefore, the stratospheric circulation responds nonlinearly to the magnitude of the precursory anomaly during the onset period of the SSW.

4 Concluding Remarks

Computations of MRI/JMA-GCM starting from the initial condition of each ensemble member for the JMA 1-month forecasts were performed to confirm the high sensitivity to the initial condition for the prediction of a WN 1 SSW event occurring in Dec 2001 reported by M05. The hindcast experiments well reproduced the similar high sensitivity to the initial condition during the onset period of the SSW.

The GCM experiments revealed that distinct U anomalies similar to those shown in M05 are significantly related to the subsequent warming in the stratospheric polar region. The characteristic tropospheric U anomalies were associated with persistent blocking over the Atlantic sector and enhanced upward propagation of WN 1 component from the troposphere into the stratosphere as in M05. Characteristic U anomalies in the stratosphere were also observed during the onset period of the SSW, which was not reported in M05. Moreover, generation of WN 1 component was much more distinct, and the statistical significance of tropospheric U anomalies was slightly reduced compared with M05. This difference suggests that the generation rather than the propagation of WN 1 is significantly related to the occurrence of the SSW. The pattern of U anomaly was also identified as the leading EOF pattern of U variations among ensemble members as in M05 during the onset period. Hence, the high sensitivity to the initial condition for the SSW prediction is connected with the coincidence between these two patterns.

The importance of the statistically obtained regressed field with respect to the subsequent SSW was also reconfirmed by conducting a series of hindcast experiments using MRI/JMA-GCM. The stratospheric warming tends to occur for initial conditions composed of ensemble mean field and the regressed field multiplied by positive coefficients although the magnitude of the warming does not linearly increase with increasing coefficient. Rather, the response is a stepwise function of the coefficient, which suggests the existence of a threshold value of the coefficient for the occurrence of the warming.

Such nonlinear feature of the stratospheric circulation reminds us the work of Yoden (1987). He showed the co-existence of two stable planetary flow regimes in the stratospheric circulation in a highly truncated atmospheric model for some parameter ranges. One corresponds to a strong polar vortex state with strong westerlies, and the other is similar to a SSW state with large-amplitude planetary waves. For such a nonlinear system, there must be a critical magnitude of initial perturbations to cause transitions from one stable state to the other across a “potential barrier” between two stable states. Moreover, there might be a preferred direction for the initial

perturbation to promote the transition at the "bifurcation point" where trajectories tend to diverge in a direction connecting these two stable states in phase space. The precursory anomalies for the SSW obtained through the regression analysis in this study would indicate such direction.

It would be possible to extract the most important ingredient for the occurrence of the SSW by conducting hindcast experiments from systematically idealized initial conditions produced from the regressed pattern. By these experiments, we could resolve the following interesting issues: the relevance of the tropospheric blocking to the subsequent SSW, the relative importance of the generation and propagation of planetary waves for the occurrence of the SSW, and the relationship between the generation of planetary waves and the tropospheric blocking. It is also very important to examine the predictability of other SSW events to reveal the robustness of the results obtained in this paper. Our preliminary analysis on the SSW event in January 2004 using JMA 1-month ensemble forecast dataset showed that the predictable period of the SSW event is at most 9 days, which is quite short compared with that of the SSW in December 2001. The SSW in December 2004 was significantly contributed by planetary wave components of smaller scales, i.e., WN 2 and 3. Therefore, it is necessary to conduct more detailed analysis to clarify the role of the smaller scale planetary waves on the predictability of the SSW events.

Acknowledgments

We are grateful to H. Yoshimura for providing us MRI/JMA-GCM. We also would like to thank all the members in Climate Prediction Divisions in the JMA for providing us 1-month forecast data sets of the JMA. This work was supported by a Grant-in-Aid for Scientific Research (B) from JSPS, and by a Grant-in-Aid for the 21st Century COE Program (Kyoto University, G3). The GFD-DENNOU Library was used for the graphics.

References

Andrews, D. G., Holton, J. R. and Leovy, C. B. (1987) *Middle Atmosphere Dynamics*. Elsevier, New York.

Arakawa, A. and Schubert, W. H. (1974) Interaction of cumulus cloud ensemble with the large-scale environment. Part I. J. Atmos., Sci., 31, 674-701.

Christiansen, B. (2003) Temporal growth and vertical propagation of perturbations in the winter atmosphere. Q. J. R. Meteor. Soc., 129, 1589-1605.

Japan Meteorological Agency (2002) Outline of the operational numerical weather prediction at the Japan Meteorological Agency (JMA). Appendix to *WMO Numerical Weather Prediction Progress Report*, JMA, Tokyo. pp. 157.

Kimoto, M., Mukougawa, H. and Yoden, S. (1992) Medium-range forecast skill variation and blocking transition: A case study. Mon. Wea. Rev., 120, 1616-1627.

Matsuno, T. (1971) A dynamical model of stratospheric sudden warming. J. Atmos., Sci., 27, 871-883.

Mechoso, C. R., Yamazaki, K., Kitoh, A. and Arakawa, A (1985) Numerical forecasts of stratospheric warming events during the winter of 1979. Mon. Wea. Rev., 113, 1015-1029.

Mechoso, C. R., Suarez, M. J., Yamazaki, K., Kitoh, A., and Arakawa, A. (1986) Numerical forecasts of tropospheric and stratospheric events during the winter of 1979: Sensitivity to the model's horizontal resolution and vertical extent. *Advances in Geophysics*, 29, 375-413.

Mellor, G. L. and Yamada, T. (1974) A hierarchy of turbulence closure models for planetary boundary layers. J. Atmos. Sci., 31, 1791-1806.

Miyakoda, K., Strickler, R. F. and Hembree, G. D. (1970) Numerical simulation of the break-down of a polar-night vortex in the stratosphere. J. Atmos. Sci., 27, 139-154.

Mizuta, R., Oouchi, K., Yoshimura, H., Noda, A., Katayama, K., Yukimoto, S., Hosaka, M., Kusunoki, S., Kawai, H. and Nakagawa M. (2006) 20-km-mech global climate simulations using JMA-GSM model. –Mean Climate States–. J. Meteor. Soc. Japan, 84, 165-185.

Mukougawa, H. and Hirooka, T. (2004) Predictability of stratospheric sudden warming: A case study for 1998/99 winter. Mon. Wea. Rev., 132, 1764-1776.

Mukougawa, H., Sakai, H. and Hirooka, T. (2005) High sensitivity to the initial condition for the prediction of stratospheric sudden warming. Geophys. Res. Lett., 32, L17806, doi:10.1029/2005GL022909.

Reichler, T., Kushner, P. J. and Polvani, L. M. (2005) The coupled stratosphere-troposphere response to impulsive forcing from the troposphere. J. Atmos. Sci., 62, 3337-3352.

Toth, Z. and Kalnay, E. (1993) Ensemble forecasting at NMC; the generation of perturbations. Bull. Am. Met. Soc., 74, 2317-2330.

Yoden, S. (1987) Bifurcation properties of a stratospheric vacillation model. J.Atmos. Sci., 44, 1723-1733.

14 Self Organized Criticality and/or Low Dimensional Chaos in Earthquake Processes: Theory and Practice in Hellenic Region

G. P. Pavlos[1], A. C. Iliopoulos[1] and M. A. Athanasiu[1]

[1] Democritus University of Thrace, Department of Electrical and Computer Engineering, 67100, Xanthi, Greece, gpavlos@ee.duth.gr

Abstract. In this study, using the modern theory of far from equilibrium nonlinear stochastic dynamics and nonlinear data analysis of seismic events in the Hellenic region, two significant aims have been accomplished: initially, the verification of the possibility for unification of two competitive and antagonistic, until now, theoretical points of view of critical complexity, such as Self Organized Criticality and Low Dimensional Chaos and subsequently, the provision of further information, about the existence of a global and low dimensional earthquake strange attractor in the Hellenic region, as it has been supported in a previous study (Pavlos et al., 1994). Furthermore, evidence for the existence of an input-output dynamical process underlying the earthquake dynamics, is found.

1 Introduction

Up to now, the Self Organized Criticality (SOC) and Low Dimensional Chaos theories are conflicted when applied for the modelling and physical understanding of random processes with power law scaling and fractal profile, as implemented in the case of earthquake processes. As it is well known, the phenomenology of earthquakes summarized in Gutenberg-Richter law, stress drop triggering, multifractality in space and time, clustering avalanches, percolation and other characteristics of seismogenesis, has been interpreted as a self organized critical process (Bak et al., 1987, 1988; Bak, 1991; Main, 1996). On the other hand, significant experimental evidence for low dimensional chaos has been reported recently in different regions of earthquake occurrence, such as Greece (Hellenic region) and Japan (Pavlos et al., 1994; Iliopoulos et al., 2007), Antarctica (Correig et al., 1997), Eastern Pyrenees (Urquizu and Correig 1998), the Caucasian region (Matcharaschvili et al., 2000), India (Srivastava et al., 1996; Tiwari et al., 2004 and Tiwari and Lakshmi, 2005), the Southern Iberian Penincula (Lana et al., 2005), and Romania (Enescu et al., 2005). Following this line of evidence concerning earthquake phenomenology and nonlinear data analysis we decided to examine the possibility of the unification of the opposite, in a first glance, concepts of SOC and Low Dimensional Chaos. As it will be shown

in the next section, significant unification can be achieved by the general critical theory of the far from equilibrium stochastic nonlinear processes. This unifying theory can produce SOC or chaotic processes and leads to a deeper understanding of the seismogenesis as well as to a constructive follow-up of nonlinear methods for modelling and forecasting.

Greek seismicity is used in this research, as well as in a series of forthcoming papers. The Greek general tectonics seismicity and fault fractality are described by Bath (1983), Makropoulos and Burton (1984), Papazachos (1990), Latoussakis et al. (1991), Papadopoulos and Dedousis (1992), Doutsos and Koukouvelas (1998) and Koukouvelas et al. (1999), Burton et al. (2004), Xu and Burton (1995, 1999, 2006).

Evidence for SOC models applied in the Greek region is given in a series of papers which examine strong earthquakes as the consequence of the progressive ordering of a local fault system (Papazachos and Papazachos, 2000; Karakaisis et al., 2002; Karakaisis, 2003). Space or time fractality, which also supports the concept of SOC for the Greek seismicity, was studied by Telesca et al. (2000, 2001, 2002), while Dimitriu et al. (2000) related Greek seismicity patterns with a percolation process.

Evidence for low dimensional chaotic seismicity in Greek region is given by Pavlos et al (1994), Plagianakos and Tzanaki (2001), Iliopoulos et al. (2007). Pavlos et al. (1994), were the first to our knowledge, to use an interevent seismic time series for testing the hypothesis of low dimensional chaos, similar to the chaotic dynamics revealed by the dripping faucet system (Shaw, 1984; Neda et al., 1996). According to this model, the random seismic events are supposed to be the consequence of lithospheric dripping through local faults.

The framework of this study is organized as follows. First, significant theoretical concepts, which are used as the fundamentals for a unifying theory of earthquake process are being introduced (Section 2) followed by nonlinear data analysis of the Greek seismicity, supporting the concept of a global earthquake strange attractor in the Greek region (Section 3). The last section consists of a summary of our results and a discussion of open problems that must be addressed to in the future.

2 The Theoretical Framework

In this section some crucial and constructive concepts, concerning the earthquake process and its relation to nonlinear dynamics are being presented, focusing on the debate concerning SOC and Low Dimensional Chaos.

2.1 The physical system

Earthquake events are caused by the deformation and sudden rupture of some parts of the Earth's crust, while the underlying dynamics is the manifestation of the global, complex and coupled Earth's system: core, mantle and crust.

The release of heat from the Earth's core causes convection in the mantle, which constitutes the driving force behind plate tectonic motions, followed by the coupling of two systems: the mantle and the lithosphere. The lithosphere moving over a weak

substrate (asthenosphere) and the mantle convection constitute the interactive physical system that creates the earthquake process (Oliver et al., 1969; Evison, 1970; Stiller et al., 1984; Doglioni, 1990; Schmelling and Marquart, 1993; Bercovici, 2002; King et al., 2002; Sornette and Pisarenko, 2003; Conrad et al., 2004).

During transient times, higher heat flow in the mantle than the respective corresponding to a steady state, can be related to time dependent or chaotic mantle convection. This shows characteristic times comparable with volcanic events (Schmelling and Marquart, 1993). The interaction between the lithosphere and the mantle convection indicates the fact that plate tectonics and lithospheric internal dynamics, related with seismogenesis, constitute an input-output driven process. According to this concept, the plate sizes and positions are driven by the internal thermal engine of the Earth that also drives mantle convection, so that the edges of moving plates against each other are sites of intense geological activity, such as earthquakes, volcanoes and mountain building (Sornette and Pisarenko, 2003).

As far as the mantle dynamics is concerned, it is well known that the thermal convection and the temperature inhomogeneities (gradient) in a spatially distributed system, can generate chaotic dynamics of the bulk motion according to Benard problem and the Landau-Ginzburg theory for systems of large spatial extent (Haken, 1983; Nicolis, 1995). On the other hand, the lithosphere is a large nonlinear complex hierarchical system of volumes of different sizes with collective behaviour and multiscale organization (Kagan, 1994; Main, 1996; Keilis-Borok et al. 2001; Sornette and Pisarenko, 2003). These characteristics clearly indicate that the lithospheric dynamics are non-autonomous and fall into the category of input-output systems (Abarbanel et al., 1993).

2.2 Seismogenesis as a nonlinear dynamic process

The nonlinear character of the earthquake dynamics is supported by a series of observational facts, as well as by nonlinear models, which mimic many spatiotemporal characteristics of seismogenesis. Some crucial and observable characteristics of nonlinear dynamics of spatially distributed systems are: scaling and power law behaviour, fractal and multifractal patterns in space and time, structured randomness (or stochasticity) discriminated by the simple infinite dimensional white noise process, critical point behaviour, avalanche dynamics and intermittency (Turcotte, 1986; Geilikman et al., 1990; Molchanov et al., 1990; Olami and Christensen, 1992; Sahimi et al., 1993; Kagan, 1994; Dongsheng et al., 1994; Davis et al., 1994; Robertson et al., 1995; Main, 1996; Bowman et al., 1998; Descherevsky et al., 2000; Telesca et al., 2000, 2001, 2002, 2004; Nakaya and Hashimoto, 2002; Kiyaschenko et al., 2004).

Burridge and Knopoff introduced the spring-block model as a simple and nonlinear model that can mimic the dynamics of earthquakes and the crack propagation along the earthquake fault (Burridge and Knopoff, 1967). The BK, nonlinear deterministic or stochastic, modelling of earthquakes has been developed in a series of studies until recently (Carlson and Langer, 1989; Nakanishi, 1990; Vieira, 1992; Vasconcelos et al., 1992; Vieira et al., 1993, 1999; Vieira and Herrmann, 1994; Vieira and Lichtenberg, 1996; Hainzl et al., 2000).

Bak et al. (1987, 1988) and Bak (1991) introduced the sandpile model, revealing characteristics of self-organized criticality, as the basic mechanism for far from equilibrium processes with spontaneous organization. This model reveals scaling laws, avalanches, intermittency and fractal spatiotemporal geometrical properties. Earthquakes constitute one of the best candidates for the SOC nonlinear mechanism as well as for avalanche type mechanisms (forest-fire models, percolation etc.) based on the SOC model (Sornette, 1992; Hooge et al., 1994; Turcotte, 1999; Turcotte et al., 1999; Newman and Turcotte, 2002).

Anghel and others introduced low-dimensional nonlinear models including chaos, describing the dynamical features of a fault zone with many degrees of freedom (Anghel et al., 2004; Anghel, 2004). Primakov and Shnirman (1999) introduced the model of movable disk with chaotic character showing significant similarity of energy dissipation process in the lithosphere and in their respective model.

Nonlinear hierarchical models for defect development and seismicity were introduced by Narkunskaya and Shnirman (1990), Blanter and Shnirman (1997), Shnirman and Blanter (1999), as well as models of renormalization group approach (Newman et al., 1994; Allegre et al., 1995; Saleur et al., 1996; Iwashita and Nakanishi, 2005).

2.3 Broad universality of earthquake models

Until now there is a controversy about the SOC attractor and the chaotic character of the earthquake process. According to SOC theory (Bak et al., 1987; Bak, 1991; Bak and Chen, 1991) the system leaves on the border of chaos. In the case of SOC attractor, the degrees of freedom are proportional to the size of the system and the uncertainty increases according to a power law. This statement is in contradiction with the chaotic exponential decay of a low dimensional attractor (Bak et al., 1988). Moreover, the SOC theory does not require an external tuning to carry the system to SOC state.

However, in a series of studies it is proved that this represents the narrow version of SOC dynamics. Yang et al. (2004) argue that earthquakes are unlikely phenomena of SOC as the first-return-time probability is changed after the rearrangement of the time series, showing that, in contrast to the SOC theory, the seismic time series is not uncorrelated. Moreover, Corral (2005) indicates that a SOC model for seismogenesis must divert from the original model of Bak-Tang-Wiesenfeld.

Vieira in a series of papers stated that SOC and strong Chaos (positive Lyapunov exponent) could coexist in the generation of Burridge and Knopoff (BK) models (Vieira, 1992; Vieira and Lichtenberg, 1996). The phenomenon of chaotic synchronization is also present in coupled subsystems in the three-block system of the BK model (Vieira, 1999). Tuning of criticality and phase transition has also been observed in definite versions of BK spring block model used to simulate the tectonic motion along the fault (Vasconcelos et al., 1992; Vieira et al., 1993). Finally, a broad universality is observed, as far as SOC and strong Chaos profiles are concerned, in any system that can mimic earthquakes (Boettcher and Paczuski, 1997; Vieira, 2004).

2.4 Broad universality of nonlinear dynamics of scaling, fractality, percolation, intermittency, renormalization, SOC, chaos and all that, in the case of seismic process

In this section a panoramic point of view of the earthquake process is presented with the use of the universal character of nonlinear dynamical processes. The following section constitutes the base for a unified picture of seismogenesis and also the base for the physical interpretation of the data results presented in the next section. For every nonlinear dynamical system of finite or infinite dimensionality (degrees of freedom) there exists a typical scenario of solutions as the system diverges from the stable state, while some external coupling (control parameter) increases. In particular, as the control parameter changes and the system moves far from equilibrium, its asymptotic motion becomes sequentially more complicated through successive bifurcations and successive appearance of low dimensional periodic attractors (limit cycles, torus), low dimensional strange attractors (chaotic solutions) and finally, for active distributed systems (continuous media), high dimensional strange attractors (state of fully developed turbulence) (Ott, 1981; Teman, 1988).

In the case of active distributed systems, the low dimensional attractors correspond to long-range spatial order (coherent patterns), while the temporal behaviour of the space coherent pattern can be chaotic. This state is known as early or few-mode turbulence (Haken, 1988; Mikhailov and Loskutov, 1991; Nicolis, 1995). It must be noted that, when macroscopical deterministic nonlinear dynamical systems are studied microscopically, they behave as nonlinear stochastic processes, described by probabilistic equations of the Fokker-Planck type. The interest in this case of nonlinear stochastic and distributed systems lies in their possibility to reveal scale invariance and long-range correlations (divergence of correlation length) at critical points (bifurcation points). This phenomenon is known as a far from equilibrium phase transition (Haken, 1983; Chang et al., 1992) and constitutes the fundamentals for a universal and unifying understanding of every intriguing phenomenon observed in earthquakes, such as: many correlation scales and scaling laws, fractal dimensions in space and time, SOC with infinite dimensionality or low dimensional chaos, intermittent turbulence and percolation states (Stinchcombe, 1988; Chang, 1992). Moreover, according to the renormalization theory applied in the case of far from equilibrium phase transition, the effective or dynamical dimensionality corresponding to the observed degrees of freedom of nonlinear stochastic distributed system depends upon the fixed points, on which the system exists and the number of relevant eigenvalues (Chang et al., 1978; Chang, 1992; Chang et al., 1992).

These characteristics of stochastic distributed nonlinear systems can be used for the interpretation of seismogenesis as a dynamical and far from equilibrium phase transition process. The term dynamical is used in order to discriminate the far from equilibrium phase transition case from the case of simple equilibrium phase transition. In the case of earthquake physical process, the nonlinear stochastic system, existing at a critical state (fixed point) and revealing the entire known observable earthquake characteristics, is a definite part of lithosphere (Main, 1996; Keilis-Borok et al., 2001; Keilis-Borok, 1990, 1999; Kossobokov, 2003; Shebalin et al., 2005). Moreover, the above

theoretical results concerning the study of nonlinear stochastic systems existing at far from equilibrium critical states are in agreement with the observed characteristics of lithosphere at regions with strong seismic activity. Particularly, fractal fault patterns and spatial earthquake distribution indicate the percolation state of lithosphere (Molchanov et al., 1990; Sahimi et al., 1992; Miltenberger et al., 1993). Also, for nonlinear stochastic systems near criticality, the correlations among the fluctuations of the random dynamic fields (stress field in the case of lithosphere) are extremely long-ranged. Moreover, under favourable conditions (strengthening of mantle and plate coupling, as the mantle convection drives plate tectonic motion and introduces free energy in the lithosphere), the state of intermittent turbulence increases, thus producing more and larger coherent structures, until some of them become global instability. In the case of lithosphere, coherent structures and instability, equals to earthquakes. According to these theoretical concepts, the development of lithospheric coherent structures and long range correlations is in accordance with the critical point hypothesis for earthquakes (Saleur et al., 1996; Bowman et al., 1998; Jaume and Sykes, 1999), as well as with the application of renormalization theory to the lithospheric dynamics, the hierarchical model of defect development, the theory of avalanches and percolation clustering (Narkunskaya and Shnirman, 1990; Lomnitz-Adler et al., 1992; Sahimi et al., 1992; Miltenberger et al., 1993; Newman et al., 1994; Allegre et al., 1995; Robertson et al.,1995; Blanter and Shnirman, 1997; Shnirman and Blanter, 1999**;** Godano and Pingue, 2005).

Finally, the development of far from equilibrium nonlinear stochastic dynamics unifies the infinite dimension SOC and low dimensional chaos processes as manifestations of the same dynamics at different fixed points, according to the renormalization group (Chang, 1992). Also, the theory of Chang verifies the results obtained by the generic type of BK spring-block modelling of earthquakes, in which the avalanche type of SOC and chaotic dynamics can coexist (Vieira and Lichtenberg, 1996; Vieira, 2004), as well as the results obtained by applications of dissipative fault modelling (Anghel, 2004).

2.5 Earthquake as an input-output driven process

As it has been stated, earthquakes, which are the bursting energy discharge manifestation of the lithosphere dynamics, are a driven process. Thus, the knowledge of input-output dynamical processes can be applied (Pristley, 1988; Broomhead and King, 1992; Abarbanel, 1993). In the case of the lithosphere, the external driver is the plate tectonics mantle convection currents, which cause a compressional stress environment. This environment acts as the stress input in the lithospheric system. Generally, the dynamics of an input-output system can be summarized in the equation

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{dt}} = \boldsymbol{f}(\boldsymbol{X}(\mathrm{t}), \boldsymbol{U}(\mathrm{t}), \lambda) \tag{1}$$

The internal state is described by the vector $\boldsymbol{X}$ and the input by the vector $\boldsymbol{U}$, while λ is the control parameter. For spatially distributed systems, $\boldsymbol{X}$ and $\boldsymbol{U}$ are fields $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{r}, \mathrm{t})$, $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{r}, \mathrm{t})$ and the dynamics can be described by partial differential equation

$$\frac{\partial \boldsymbol{X}}{\partial \mathrm{t}} = \boldsymbol{f}(\{\boldsymbol{X}(\boldsymbol{r},\mathrm{t}),\boldsymbol{U}(\boldsymbol{r},\mathrm{t})\},\nabla^{\mathrm{k}}\{\boldsymbol{X}(\boldsymbol{r},\mathrm{t}),\boldsymbol{U}(\boldsymbol{r},\mathrm{t}),\lambda(\boldsymbol{r},\mathrm{t})\}) \quad (2)$$

Subsequently, in an input-output process the internal dynamical system is coupled with an environment, while providing an output corresponding to the observed signal *z(t)*. The output signal includes information of the internal dynamical process, as well as of the input signal $\boldsymbol{U(t)}$, according to the relation

$$\mathrm{z}(\boldsymbol{r},\mathrm{t}) = \Phi(\boldsymbol{X}(\boldsymbol{r},\mathrm{t}),\boldsymbol{U}(\boldsymbol{r},\mathrm{t})) \quad (3)$$

In the case of nonlinear systems, the internal dynamics can be of different types: stable, limit point, periodical with discrete spectrum (limits cycle, limit torus) or chaotic, corresponding to strange attractors.

As it is shown by Pavlos et al (1999) and Athanasiu and Pavlos (2001), time series produced by input-output dynamics can be studied by an extended algorithm of nonlinear analysis which is able to discriminate crucial characteristics of the external coupling, as well as characteristics of the internal dynamics. Summarizing the results of the pre-mentioned studies, it is useful, concerning the lithosphere-mantle coupling, to state that:

a) The correlation dimension estimated by using the slopes of the correlation integrals corresponding to the observed signal give valid information, not only concerning the dimensionality of the external input $U(t)$ but also for the dimensionality of the internal dynamics.
b) Singular Value Decomposition (SVD) analysis also gives solid information about the existence of external coupling corresponding to deterministic signal or to white-colored noise.

Generally, when the internal system is spatially distributed, the dimensionality of the output signal is practically infinite. However, far from equilibrium, the infinite degrees of freedom can be reduced to a finite number, by a self-organizing process, which creates patterns of long-range spatial order and coherent patterns with periodic or chaotic temporal behavior (few-mode turbulence). This case of few-mode turbulence can be described as the existence of a global finite dimensional attractor. Furthermore, when the control parameter λ, is changed, the effective dimensionality of the few mode turbulence may increase or/and the system can make phase transition to the state of developed turbulence. In the latter case, it is possible for independent local low dimensional attractors to exist (Lorenz, 1991; Tsonis, 1992). According to the above concepts, we can hypothesize that it is possible for local regions of the lithosphere to reveal low dimensional dynamics. As far as the input signal is concerned, it may also be produced by a low or high (practically infinite) dimensional process. As it was shown by Pavlos et al. (1999), the slopes of the correlation integral of the output signal, in many cases, apart from the internal dimensionality, can also reveal the dimensionality of the input signal. Furthermore, SVD analysis can strongly indicate the existence of external dynamics (Elsner and Tsonis, 1996; Pavlos et al., 1999; Athanasiu and Pavlos, 2001).

2.6 SOC, chaos and earthquake clustering

Earthquakes display a complex spatiotemporal behavior as they occur in clusters and reveal scale invariant patterns and power law correlations in space and time. Some of the patterns, as recognized by seismologists, are foreshocks, aftershocks, precursory activation and quiescence. Although there is a considerable stochastic component in the earthquake sequence, which is the sequence of main shocks (close Poissonian), the concept of seismic cycle (Fedotov, 1968) postulates a high degree of regularity in location, magnitude and time of earthquake occurrence. According to the concept of seismic cycle, there are three stages of seismogenesis: the aftershock sequence of the earlier mainshock, a very long interval of comparatively low seismicity and an accelerating activity, which culminates in the later mainshock, often called the stage of accelerated moment release or precursory stage (Fedotov, 1968; Karakaisis et al., 1991; Bufe and Varnes, 1993; Jaume and Sykes, 1999; Evison, 2001).

The concept of seismic cycle indicates the holistic character of the seismic process, in contrast to the older view of two separated physical processes or relaxation mechanisms, one for the main shocks, which ought to follow an exponential law (Poisson distribution) and an independent process, which follow a power law (Omori's law). This reductionistic view has been refuted in a series of studies (Bak et al., 2002; Davidsen and Goltz, 2004; Lindman et. al., 2005; Saichev and Sornette, 2006; Corral, 2003, 2004, 2005, 2006, 2006, 2006; Shcherbakov et al., 2005, 2006), in which all the earthquake events are placed on the same footing, whether they are classified as main shocks or aftershocks. Thus, we can accept the hypothesis that earthquakes, during a seismic cycle, constitute a whole process which includes correlated events (foreshocks, aftershocks) as well as uncorrelated (Poissonian) events (main shocks).

On the other hand, the existence of power law correlation of aftershock events was used as an argument that earthquakes are unlikely phenomena of SOC (Yang et al., 2004). According to this point of view if earthquakes are indeed caused by a SOC mechanism, they cannot be discriminated in mainshocks and aftershocks, as they occur randomly anywhere and in any time and an earthquake can not know how large it will become. However, this statement is in contrast with observations which reveal power law distributions for the first return-time probability $P_M(T)$. According to Corral (2005), the identification of SOC with the total absence of correlations and the failure to reproduce the occurrence of foreshock and aftershock sequences correlated with large earthquakes, was a weakness of the first generation of SOC earthquake models (Burridge and Knopoff, 1967; Bak, Tang and Wiesenfeld, 1987; Bak and Tang, 1989), which display an exponential distribution of recurrence times and fail to reproduce the spatiotemporal clustering of smaller events accompanying large earthquakes.

In this direction, the efficiency of the SOC type earthquake models to exhibit sequences of foreshocks and aftershocks with power-law scaling is supported in a series of studies (Olami and Christensen, 1992; Olami, Feder and Christensen, 1992; Hainzl et al., 2000; Davidsen and Schuster, 2000; Hergarten and Neugebauer, 2002; Helmstetter et al., 2004; Weatherley, 2006; Woodard et al., 2006; Peixoto and Prado,

2006). Moreover, the complex (holistic) and hierarchical character of the seismic cycle is supported by the general properties of earthquake sequences in real catalogs where: the aftershocks can also produce their own aftershock series or swarms or series of earthquakes that are impossible to be divided into foreshocks, main shocks and aftershocks, as they have similar magnitudes (Gabrielov et al., 1990). Also, an entire mainshock-aftershock event can be nested inside a larger aftershock sequence (Utsu, 1970). As the seismic cycle is not a cycle, in the usual sense of the word, periods between the strongest earthquakes are not equal and deviate considerably from the average characteristic period. The above experimental characteristics of seismogenesis further support the hypothesis of a local strange attractor underlying the seismic process (Gabrielov et al., 1990; Pavlos et al., 1994; Sykes et al., 1999; Evison, 2001; Keilis-Borok et al., 2001; Turcotte, 2001) causing short time correlations, foreshock and aftershock events, and long time decorrelations of (Poissonian) mainshock seismic events. This hypothesis is in accordance with the theoretical concepts introduced in the previous sections of this study.

3 Data Analysis and Results

In this section significant results applying nonlinear analysis at seismic time series are being presented. The time series correspond to 11937 seismic events that have occurred in the Hellenic region (Greece), with magnitude greater than 3.2 Richter and were observed during the period of 1968-1993. The time series was constructed by the bulletins of the National Observatory of Athens. Following the concept that the earthquakes fall into the category of dripping process (Pavlos et al., 1994), we have created a time series of time intervals, between successive earthquakes.

3.1 Randomness, determinism, low dimensionality, nonlinearity and chaoticity

Figure 14.1 presents the time series corresponding to the interevent time intervals (Fig. 14.1a), the autocorrelation coefficient (Fig. 14.1b) and the amplitude distribution of the entire signal as well as its first and second half (Fig. 14.1c). In these figures we can observe that the signal decorrelates after twenty-thirty lag times, and although it is probably produced by a nonlinear random process, it is stationary. However, while the decaying autocorrelation function indicates the random character of the signal, it does not provide any information about the kind of randomness, except that it is not a form of white noise, as the decorrelation time of white noise is zero.

In order to exclude the case of the high dimensional fractal Brownian motion (colored noise) we have estimated the structure function

$$S(n) = \sum_{i=1}^{N-n} [x(t_i + n\Delta t) - x(t_i)]^2 \tag{4}$$

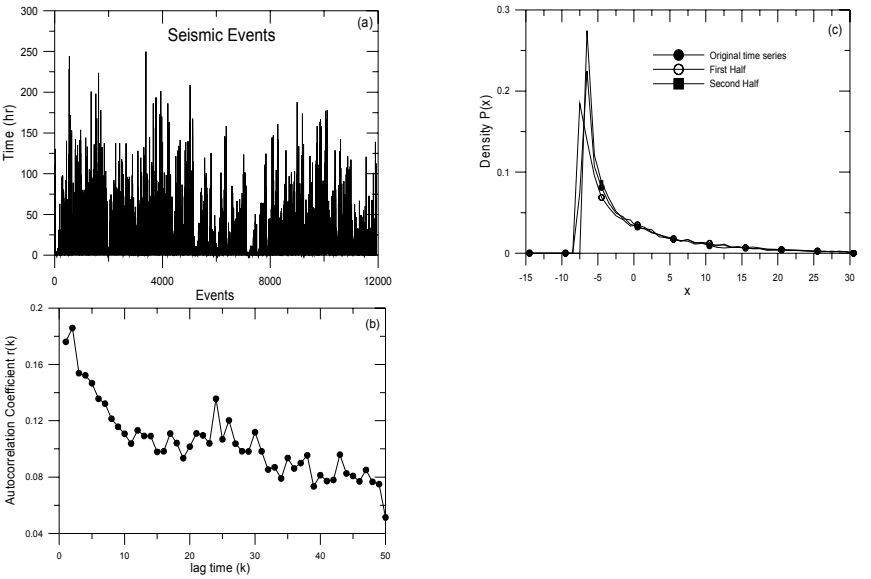


Fig. 14.1. (a) Time series of time differences between sequentially earthquake events with magnitude over 3.2 Richter, occurred in Greece. (b) The autocorrelation coefficient for the first 50 units of lag time. (c) Amplitude distribution of the original time series and its first and second half.

of the signal according to Provanzale (1991). For the case of a fractal signal the structure function reveals a scaling behavior $S(n) \sim n^{2H}$, where H is related to the fractal dimension $D = 2 - H$. The above scaling relation is related to the self-similar character of fractal Brownian motion (Takayasu, 1990). Figure 14.2 represents the structure function as a function of n. As it can be observed, the structure function reveals saturation as n is increased. This indicates that the dynamics are finite dimensional, while the region of the state space visited by the underlying dynamical system is limited.

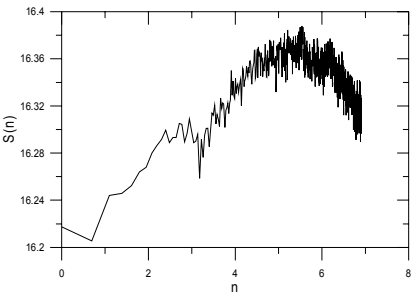


Fig. 14.2. A log-log plot of the structure function *S(n)* as a function of *n*, for the original time series.

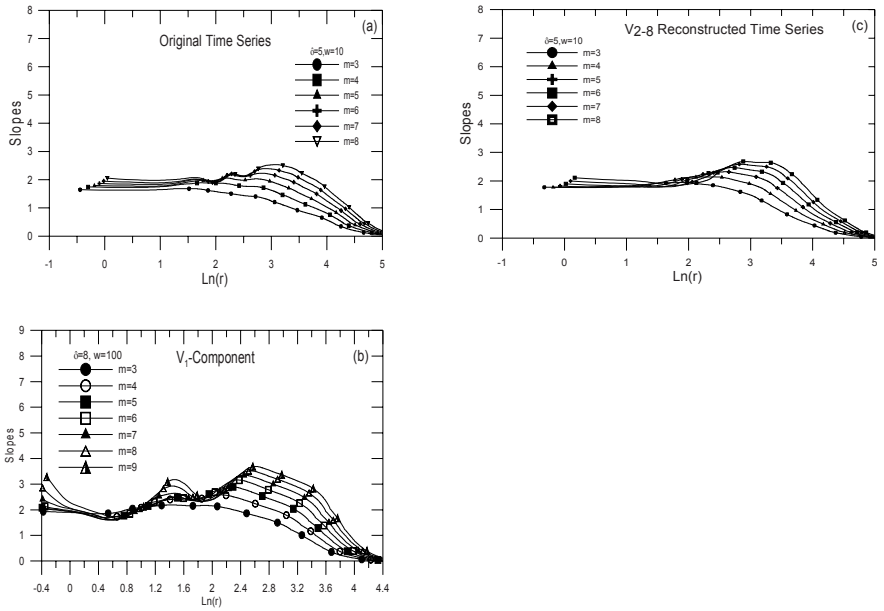


Fig. 14.3. (a –c) The slopes of the correlation integrals estimated for the original time series and its V_1 and $V_{2\text{-}8}$ SVD components, for delay time τ = 5 and Theiler parameter w =10, 100, 10 and for embedding dimensions m=3 –8, as a function of *ln(r)*.

Figure 14.3(a-c) represents the slopes of the correlation integrals estimated using the algorithm of Grassberger and Procaccia (1983), for embedding dimension, m=3-8, for the original seismic intervened time series, as well as the respective SVD components V_1 and the reconstructed times series $V_{2\text{-}8}$, according to Broomhead and King (1986), Elsner and Tsonis (1996), Iliopoulos et al. (2007).

In order to exclude the case of low value saturation of the slopes caused by the self-affinity or "crinkliness" of the trajectory in the state space, we have omitted time correlated state points, according to Theiler (1994). The low value saturation D≈2±0.25 of the slopes reveals that the underlying dynamics of earthquake time series correspond to a low dimensional deterministic process. According to the embedding theory (Takens, 1981) the dynamical degrees of freedom can not be higher than $2D + 1 \approx 5$ and lower than $D_{\text{int}} \approx 3$ where D_{int} is the first integer value after the saturation value $D \approx 2$.

However, this conclusion of low value dimensionality does not strictly imply nonlinearity of the underlying process, due to the fact that according to Theiler et al (1992a, b) the observed signal could be caused by a nonlinear monotonic static distortion of a stochastic white noise signal. In order to exclude the case of an infinite dimensional stochastic process that can mimic the profile of low dimensional process as that which is supposed for the observed time series (null hypothesis), the method

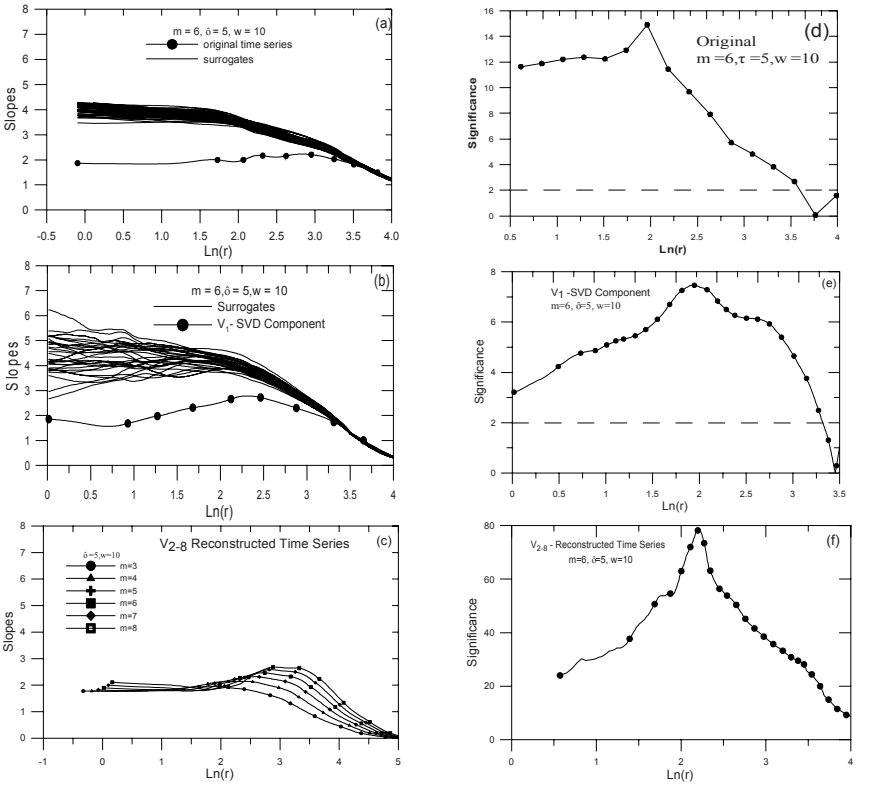


Fig. 14.4. (a) Slopes of the correlation integrals estimated for the original signal and its *30* surrogate data as function of *ln(r).* (b-c) Slopes of the correlation integral estimated for the SVD components of the original signal, V_1 and $V_{2\text{-}8}$, and their *30* surrogate data as function of *ln(r).* (d–f) The significance of the discriminating statistics as function of *ln(r)*, shown in (a-c), respectively.

of Theiler (1992a, b) was applied. Thus, an efficient number of surrogate time series was constructed, which have the same autocorrelation function and probability density with the observed time series, using the improved algorithm of Schreiber and Schmitz (1996). When the value of the significance of the discriminating statistics is higher than 2-3, then the probability that the observed seismic signal does not belong to the same family of the respective surrogate data is higher than 0.95-0.99. According to this line of evidence, Fig. 14.4 (a-c) represents the slopes of the correlation integral estimated for embedding dimension m=6 and for the three cases of Fig. 14.3 (original time series and its SVD components), as well as the slopes of efficient number of surrogate time series. In the same figure the significance of the discriminating statistics for the three cases is presented (Fig. 14.4d-f). The values of significance, which appear to be much higher than two, clearly demonstrate nonlinear and deterministic process. In this figure we can also observe the fact that although the V_1 component is safely discriminated from the case of linear stochasticity, nevertheless

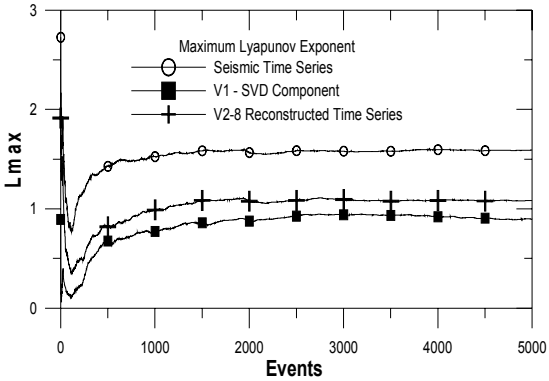


Fig. 14.5. The largest Lyapunov exponent L_{max} as a function of the seismic events, for the seismic time series and its SVD components.

it reveals the smaller statistics values of the significance. As it is further supported, this noticeable difference of V_1 compared to the next components, indicates the existence of two independent but coupled seismic processes.

Until now the obtained information concerning the underlying process of the studied seismic signal strongly supports the low dimensional (N =3-5) as well as the nonlinear character of the corresponding earthquake dynamics. Although the random character of the signal, obtained by its fast decorrelation profile, strongly supports the chaotic character of the nonlinear dynamics, in order to exclude any other cause of randomness we have estimated the Largest Lyapunov Exponent (Lmax) for the original signal and its SVD components, as shown in Fig. 14.5. The Lmax estimation was achieved according to Wolf et al (1985). As it is shown in Fig. 14.5, Lmax is clearly positive for the original signal, as well as for its SVD components. This proves the existence of sensitivity to initial conditions and thus the chaotic character of the earthquake dynamics. Also, in the same figure it can be observed that the lowest Lmax value was obtained for the V_1 component. This result also indicates the possibility of the existence of two discriminated and coupled dynamics underlying the seismic process, which is in accordance with the previous interpretation of Fig. 14.4(b, e). If this hypothesis stands correct then the $V_{2\text{-}8}$ components correspond to the more chaotic process (higher values of Lmax) and V_1 corresponds to a more stable but also chaotic process.

3.2 Singular value spectrum and input-output dynamics

In order to test this hypothesis, the spectrum of singular values σ_i corresponding to the SVD components of the original signal has been estimated. According to Broomhead and King (1986) and to Elsner and Tsonis (1986), the deterministic components correspond to the ordering $\sigma_1 > \sigma_2 > \ldots > \sigma_k$, of the first k values, while the next singular values correspond to the noise components saturating at noise level, where σ_n = const., for $n > k$. Figure 14.6 represents the singular value spectrum for the original

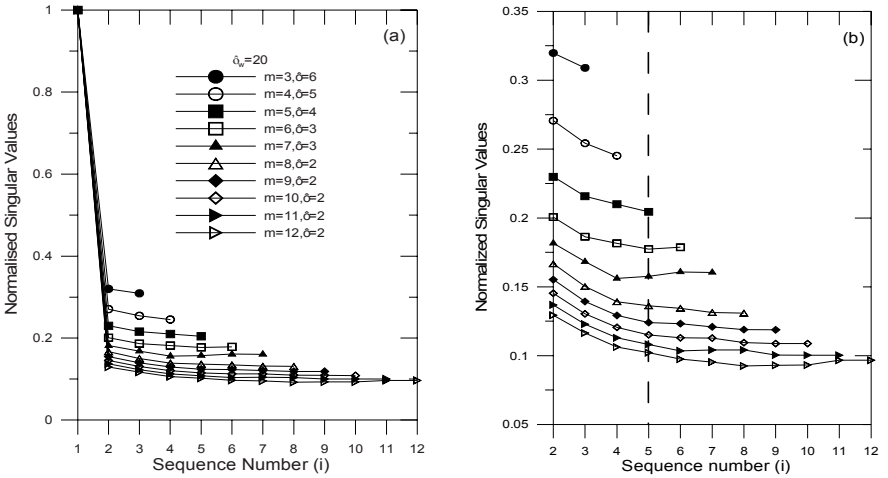


Fig. 14.6.(a) The spectrum of normalized singular values σ_i, $i = 1\text{-}8$. (b) The same with (a) but discarding the first singular value.

time series. Figure 14.6a represents the first twelve singular values and Fig. 14.6b, all singular values excluding the first. As we can conclude based on this figure, the noise level is obtained after the first five (5) singular values, indicating that the number N of the deterministic components is $N \approx 5$. This value corresponds to the effective degrees of freedom and is in agreement with the value $N=2D+1 \approx 5$, estimated previously by the slopes of the correlation integrals. However, the same figure shows that the first singular value σ_1 corresponding to the V_1 SVD component compresses the next deterministic singular values to the noise level. This fact clearly demonstrates the existence of two coupled dynamics, one external related to the V_1 component and the second internal, related to the next SVD components.

In order to further support this point of view, the spectrum of singular values estimated for the purely deterministic Lorenz system and a stochastic Lorenz system is presented in Fig. 14.7a. The stochasticity was caused by external coupling with white noise component. In the same figure, the singular values estimated for the SVD components V_1 and $V_{2\text{-}15}$ of the stochastic signal are also shown. This figure demonstrates that the perturbation of the dynamical system by white noise leaves unchanged the nontrivial eigenvalues in all cases: deterministic Lorenz system, stochastic Lorenz system and their SVD components.

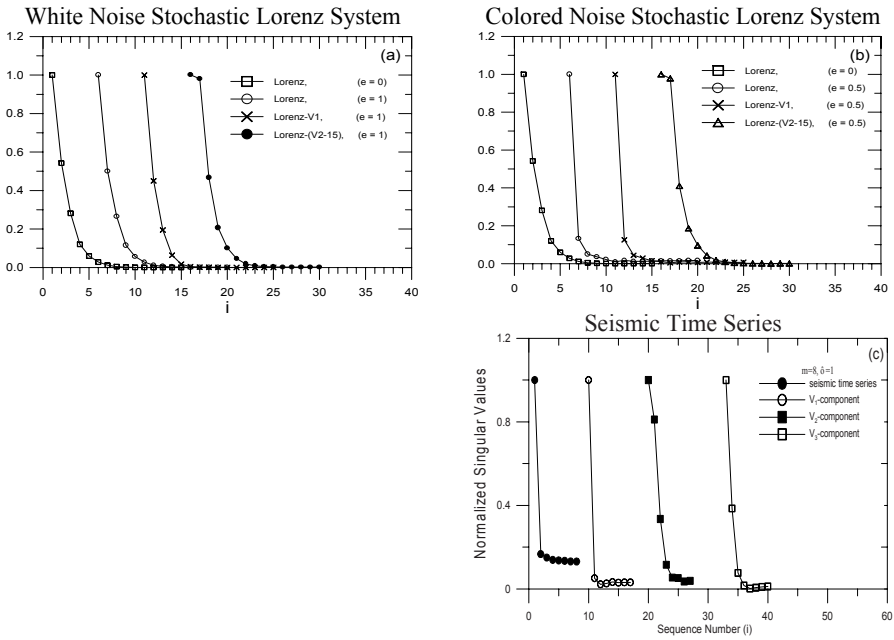


Fig. 14.7. a, b) The spectra of the singular values σ_i , *i = 1-25* corresponding to the purely deterministic *x(t)* Lorenz signal, the stochastic *x(t)* Lorenz signal and its SVD components estimated for the two noise levels *e = 1 (7.8%)* and *e =0.5 (185%)* of external white and colored noise perturbation, respectively. c) The spectra of the singular values σ_i , *i = 1-8* corresponding to the original time series and its V_1, V_2, V_3 SVD components estimated for m=8 and τ=1.

Figure 14.7b provides the same information as Fig. 14.7a, except from the fact that stochasticity was caused by external colored noise component. As it is observed in this figure, the coupling with the colored noise signal destroys the normal character of the nontrivial eigenvalues only in the case of the stochastic signal and its first (V_1) SVD component. However, this is clearly not the case in the next SVD components, where the spectrum returns to its original profile. This result is expected, as the external coloured noise dynamics exist on smaller frequencies, which are related also to the first SVD component. Due to this synchronization, the first singular value compresses the next ones to the noise floor. Fig. 14.7c represents the spectra of the singular values estimated for the original seismic time series and its SVD components. As it can be observed, the first singular value compresses the next ones to the noise floor, for both cases of the original signal and its V_1 SVD component. Here clearly we observe an obvious similarity with the case of Lorenz system coupled with the external coloured noise (Fig. 14.7b). This indicates, as far as the seismic process is concerned, the existence of an external coupling. However, although the coloured noise is a high dimensional process, it is possible for a low dimensional external coupling to exist, which can also suppress the singular value spectrum.

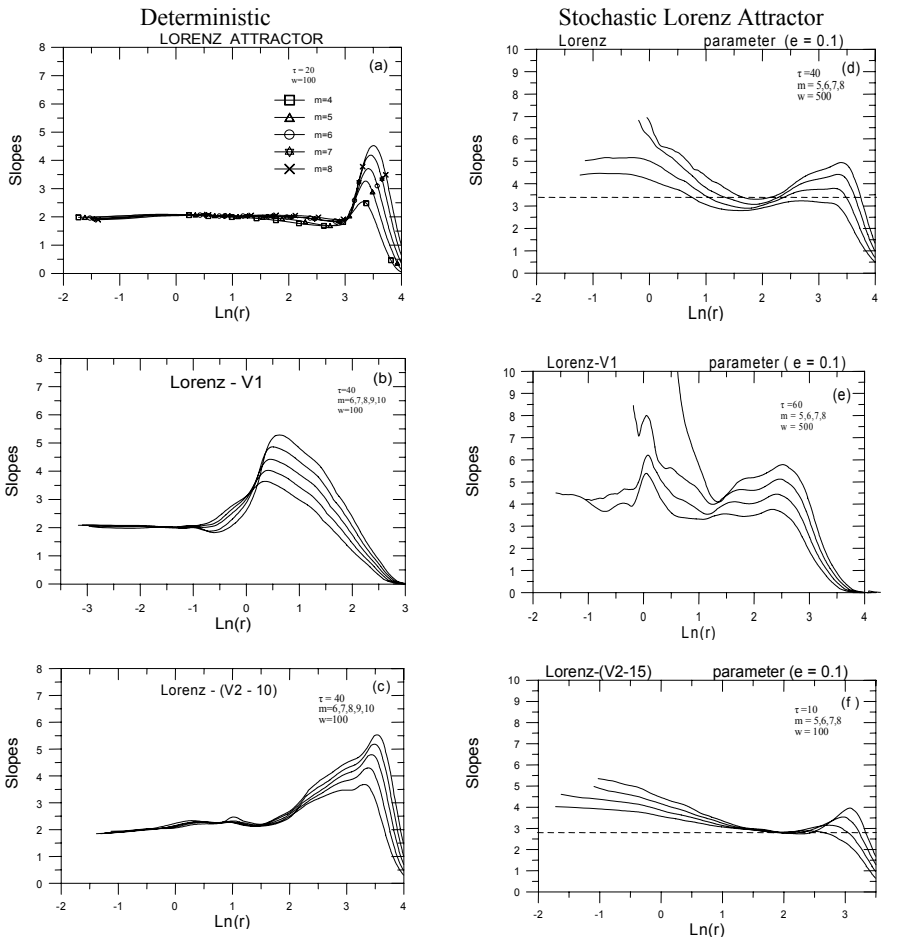


Fig. 14.8. (a-c) The slopes of the correlation integrals estimated for the purely deterministic *x(t)* Lorenz signal and its SVD components V_1 and $V_{2\text{-}15}$. (d-f) The slopes of the correlation integrals estimated for the colored noise stochastic Lorenz time series, *e =0.1 (37%),* and its SVD reconstructed components, as a function of *ln(r)*.

Thus, in order to decide about the dimensionality of the external component, we proceed to comparison of the slopes of the correlation integrals of the original time series and its SVD components, in both cases, for the purely deterministic and for the coloured noise perturbed Lorenz system, as shown in Fig. 14.8. For the purely deterministic Lorenz system the slope profile remains invariant for the original time series and its SVD components (Fig. 14.8a-c). However, the coloured noise perturbation clearly modified the slope profile (Fig. 14.8d-f). That is in the case of original stochastic signal, we distinguish only a small trace of saturation at the value of

$D \sim 3$-5, without scaling profile (Fig. 14.8d). In the case of the first (V_1) SVD component there is no low value saturation (Fig. 14.8e). However, for the higher SVD components, the slope profile clearly reverts to near the original state of low values (Fig. 14.8f). On the contrary, for the case of the seismic time series (Fig. 14.3), the profile of the slopes remains invariant for the original signal and its SVD components. The last point indicates some discrepancy with our previous hypothesis about the existence of external coupling of the earthquake process according to the singular value profile (Fig. 14.7b-c). This discrepancy can be surpassed if in the case of earthquake process we proceed to two suppositions: first that the external component is low dimensional and second that there exists a kind of synchronization between the two coupled dynamics (Abarbanel et al., 1993; Vieira, 1999). Thus, it can be hypothesized, that a process of synchronization creates identical saturation value of the slopes of the external and internal dynamics making impossible the discrimination of the two independent and coupled processes. Finally, the profile of singular value spectra, surrogate data and the Lmax exponent, as been studied for the seismic time series and its SVD components, strongly supports our hypothesis about two, coexisted and coupled dynamics, according to the input-output picture.

4 Summary and Discussion

This study represents an enterprise for the unification of different and conflicting points of view about seismogenesis by using modern theoretical concepts, as well as results of nonlinear data analysis concerning earthquakes obtained in the Greek (Hellenic) region. At the theoretical level, the concept of SOC and Chaos unification has been introduced, based on the general theory of far from equilibrium nonlinear stochastic systems. For these systems, the renormalization theory produces the spectrum of possible fixed points corresponding to SOC state or Chaos state, on which the far from equilibrium critical dynamics can exist. Thus, the application of renormalization group theory at far from equilibrium critical dynamics includes the possibility of scaling laws along with long-ranged and multiscaled correlations of fluctuations near criticality, which is in agreement with the original SOC theory, as well as the possibility of low dimensional chaos, when the number of surviving positive eigenvalues (known in renormalization theory as relevant eigenvalues) are small.

As for seismic data observed in Hellenic region are concerned, we studied the underlying dynamics in the reconstructed state space corresponding to the interevent seismic time series, according to the dripping faucet model applied for seismogenesis. Particularly, we estimated the singular value spectrum, the correlation integrals and their slopes and the maximum Lyapunov exponent, for the original seismic time series and its SVD components. In order to exclude the case of linear stochastic dynamics that mimic low dimensional chaos, we have compared our results with an efficient number of surrogate data. Furthermore, we used the Lorenz system perturbed by external signals as an explicative input-output paradigm of the profile of the slopes, the singular value spectra and Lmax of the seismic data. The results of the above analysis gave us solid information, so as to support the following:

a) The existence of low dimensional global strange earthquake attractor corresponding to the Greek lithospheric dynamics.
b) The possible existence of an external low dimensional and maybe chaotic driver. This driver concerns the stress environment of the Greek lithospheric region which corresponds to the dynamics of the mantle convection
c) The existence of a type of dynamical synchronization of the external driver (mantle convection) and the driven Greek lithospheric region.

It remains in the future to make use of this information about the existence of a global earthquake attractor for nonlinear modelling and prediction of seismicity. Moreover, it is imperative to further examine the seismicity of the Greek region in order to test the hypothesis of the interplay of the global earthquake strange attractor, supported in this paper, with the possibility of the existence of local earthquake attractors, which can work synergistically. This hypothesis may also demonstrate a pattern for modelling and prediction of the Greek seismicity as a synthesis of local and synchronized chaotic dynamics.

Acknowledgments

We would like to thank Prof. A. A. Tsonis for having kindly and constructively lead Prof. G. P. Pavlos and his co-workers for the last fifteen years, in the tools and methods of nonlinear data analysis. We would also like to thank one of the reviewers for helping us to take a deeper look into the meaning of the results of our analysis.

References

Abarbanel, H. D., Brown, R., Sidorowich, J. J., Tsirming, L. S. (1993) The analysis of observed chaotic data in physical systems. Rev. Mod. Phys. 65, 1331-1392.

Allegre, C. J., Mouel, J. L. L., Chau, H. D., Narteau, C. (1995) Scaling organization of fracture tectonics and earthquake mechanism. Phys. Earth Planet. Int. 92, 215-233.

Anghel, M. (2004) On the effective dimension and dynamic complexity of earthquake faults. Chaos, Solitons and Fractals 19, 399-420.

Anghel, M., Ben-Zion, Y. and Rico-Martinez, R. (2004) Dynamical System Analysis and forecasting of deformation produced by an earthquake fault. Pure Appl. Geophys. 161, 2023-2051.

Athanasiu, M. A. and Pavlos, G. P. (2001) SVD analysis of the Magnetospheric AE index time series and comparison with low dimensional chaotic dynamics. Nonlin. Proces. Geophys. 8, 95-125.

Bak, P., Tang, C. and Wiesenfeld, K. (1987) Self-Organized Criticality: An Explanation of 1/f Noise. Phys. Rev. Lett. 59(4), 381-384.

Bak, P., Tang, C. and Wiesenfeld, K. (1988) Self-Organized Criticality. Phys. Rev. A 38(1), 364-374.

Bak, P. and Tang, C. (1989) Earthquakes as Self-Organized Critical Phenomenon. J. Geophys. Res. 94(B11), 15635-15637.

Bak, P. and Chen, K. (1991) Self-organized criticality. Scientific American, 26-33.

Bak, P. (1991) Catastrophes and Self-Organized Criticality. Computers in Physics , 430-433.
Bak, P., Christensen, K., Danon, L. and Scanlon, T. (2002) Unified Scaling Law for Earthquakes. Phys. Rev. Lett. 88(17), 178501(4).
Bath, M. (1983) The seismology of Greece. Tectonophysics 98, 165-208.
Bufe, C. G., Varnes, D. J. (1993) Predictive modelling of the seismic cycle of the Greater San Francisco Bay region. J. Geophys. Res. 98, 9871-9883.
Blanter, E. M. and Shnirman, M. G. (1997) Simple hierarchical systems: stability self-organized criticality and catastrophic behaviour. Phys. Rev. E 55(6), 6397-6403.
Bercovici, D. (2002) The generation of plate tectonics from mantle convection. Earth Plan. Sci. Lett. 205, 107-121.
Boettcher, S. and Paczuski, M. (1997) Broad universality in self-organized critical phenomena. Physica D 107, 173-173.
Bowman, D. D., Ouillon, G., Sammis, C. G., Sornette, A. and Sornette, D. (1998) An observational test of the critical earthquake concept. J. Geophys. Res. 103(B10), 24359-24372.
Broomhead, D. S. and King, G. P., (1986) Extracting qualitive dynamics from experimental data, Physica D, 20, 217-236.
Burridge, R. and Knopoff, L. (1967) Model and theoretical seismicity. Bull. Seismol. Soc. Am. 57, 341.
Burton, P. W., Xu, Y., Qin, C., Tselentis, G-Akis, Sokos, E. (2004) A catalogue of seismicity in Greece and the adjacent areas for the twentieth century. Tectonophysics 390, 117-127.
Carlson, J. M., and Langer, J. S. (1989) Properties of the Earthquakes generated by Fault dynamics. Phys. Rev. Lett. 62, 2632-2635.
Chang, T. (1992) Low-Dimensional Behavior and Symmetry Braking of Stochastic Systems near Criticality – Can these Effects be Observed in Space and in the Laboratory. IEEE 20(6), 691-694.
Chang, T., Nicoll, J. F. and Young, J. E. (1978) A closed-form differential renormalization group generator for critical dynamics. Physics Letters 67A(4), 287-290.
Chang, T., VVedensky, D. D. and Nicoll, J. F. (1992) Differential Renormalization Group Generators for static and dynamic critical phenomena. Physics Reports 217(6), 279-360.
Conrad, C. P., Bilek, S. and Lithgow-Bertelloni, C. (2004) Great earthquakes and slab pull: interaction between seismic coupling and plate-slab coupling. Earth Plan. Sci. Lett. 218, 109-122.
Corral, A. (2003) Local distributions and rate fluctuations in a unified scaling law for earthquakes. Phys. Rev. E 68(3), 035102(4).
Corral, A. (2004) Long-Term Clustering, Scaling, and Universality in the Temporal Occurence of Earthquakes. Phys. Rev. Lett. 92(10), 108501(4).
Corral, A. (2005) Time-decreasing hazard and increasing time until the next earthquake. Phys. Rev. E 71(1), 017101(4).
Corral, A. (2005) Comment on "Do Earthquakes Exhibit Self-organized Criticality", Phys. Rec. Lett., 95, 159801.
Corral, A. (2006) Dependence of earthquake recurrence times and independence of magnitudes on seismicity history. Tectonophysics 424, 177-193.
Corral, A. and Christensen, K. (2006) Comment on "Earthquakes Descaled: On Waiting Time Distributions and Scaling Laws". Phys. Rev. Lett. 96 (10), 109801 (1).
Corral, A. (2006) Universal Earthquake-Occurrence Jumps, Correlations with Time and Anomalous Diffusion. Phys. Rev. Lett. 97(17), 178501(4).
Correig, A. M., Urquizu, M., Vila, J. and Marti, J. (1997) Analysis of the Temporal Occurrence of Seismicity at Deception Island (Antarctica), A Nonlinear Approach. Pure Appl. Geophys. 149, 553-574.

Davidsen, J. and Goltz, C. (2004) Are seismic waiting time distributions universal?. Geophys. Res. Lett. 31, L21612(4).

Davis, A., Marshak, A., Wiscombe, W. and Cahalan, R. (1994) Multifractal characterizations of nonstationarity and intermittency in geophysical fields: Observed retrieved or simulated. J. Geophys. Res. 99(D4), 8055-8072.

Descherevsky, A. V., Lukk, A. A. and Sidorin, A. Ya. (2000) Evidences of Self-Organization in Geophysical Fields' temporal variations. Phys. Chem. Earth (A) 25(12), 775-779.

Dimitriu, P. P., Scordilis, E. M. and Karakostas, V. G. (2000) Multifractal Analysis of the Arnea Greece Seismicity with Potential Implications for Earthquake Prediction. Natural Hazards 21, 277-295.

Doglioni, C. (1990) The Global Tectonic Pattern, J. Geodynamics 12, 21-38.

Dongsheng L., Zhaobi, Z., Binghong, W. (1994) Research into the multifractal of earthquake spatial distribution. Tectonophysics 233, 91-97.

Doutsos, T. and Koukouvelas, I. (1998) Fractal analysis of normal faults in Northwestern Aegean Area Greece. J. Geodynamics 26(2-4), 197-216.

Elsner J., B., and Tsonis, A. A. (1996) *Singular spectrum analysis, a new tool in time series analysis*. Plenum Press, New York.

Enescu, B., Ito, K., Radulian, M., Popescu, E. and Bazacliu, O. (2005) Multifractal and Chaotic Analysis of Vrancea (Romania) Intermediate-depth Earthquakes: Investigation of the Temporal Distribution of Events. Pure Appl. Geophys. 162, 249-271.

Evison, F. F. (1970) Seismogenesis. Tectonophysics 9, 113-128.

Evison, F. F. (2001) Long-range synoptic earthquake forecasting: an aim for the millennium. Tectonophysics 338, 207-215.

Fedotov, S.A. (1968) The seismic cycle, quantitative seismic zoning, and long term seismic forecasting. In: Medvedev, S.V. (Ed.), *Seismic Zoning in the USSR*. Izdatel'stvo Nauka, Moscow, 133-166.

Gabrielov, A. M., Levshina, T.A. and Rotwain, I.M. (1990) Block model of earthquake sequence. Phys. Earth. Planet. Int. 61, 18-28.

Geilikman, M. B., Golubeva, T. V., Pisarenko, V. F. (1990) Multifractal patterns of seismicity. Earth Plan. Sci. Lett. 99, 127-132.

Godano, C., and Pingue, F., (2005) Multiscaling in earthquakes diffusion. Geophys. Res. Lett. 32, L18302, doi:10.1029.

Grassberger, P. and Procaccia, I. (1983) Measuring the strangeness of strange attractors. Physica D 9, 189-208.

Hainzl, S., Zoller, G. and Kurths, J. (2000) Self-organization of spatio-temporal earthquake clusters. Nonlin. Proces. Geophys. 7, 21-29.

Haken, H. (1983) *Advanced Synergetics*, Springer-Verlag, New York.

Helmstetter, A., Hergarten, S., Dornette, D., (2004) Properties of foreshocks and aftershocks of the nonconservative self-organized critical Olami-Feder-Christensen model. Phys. Rev. E 70(4), 046120(13).

Hergarten, S. and Neugebauer, H. J. (2002) Foreshocks and Aftershocks in the Olami-Feder-Christensen Model. Phys. Rev. Lett. 88(23), 238501(4).

Hooge, C., Lovejoy, S., Schertzer, D., Pecknold, S., Malouin, J. F. and Scmitt, F. (1994) Multifractal phase transitions: the origin of self organized criticality in earthquakes. Nonlin. Proces. Geophys. 1, 191-197.

Iliopoulos, A. C., Athanasiu, M. A. and Pavlos, G. P. (2007), Nonlinear Analysis of Seismic data in Hellenic region and strong evidence for low dimensional chaos. (submitted for publication)

Iwashita, Y., Nakanishi, I. (2005) Scaling laws of earthquakes derived by renormalization group method. Chaos, Solitons and Fractals 24, 511-518.

Jaume, S. C., and Sykes, L. R., (1999) Evolving towards a Critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. Pure appl. Geophys. 155, 279-306.

Kagan, Y. Y. (1994) Observational evidence for earthquake as a nonlinear dynamic process. Physica D 77, 160-192.

Karakaisis, G. F., Kourouzidis, M. C., Papazachos, B. C. (1991) Behaviour of seismic activity during a single seismic cycle. Earthquake Prediction: State of the Art, Strasbourg, France, 15-18 October, 47-54.

Karakaisis, G. F., Papazachos, C. B., Savvaidis, A. S. and Papazachos, B. C. (2002) Accelerating seismic crustal deformation in the North Aegean Trough Greece. Geophys. J. Int. 148, 193-200.

Karakaisis, G. F (2003) Accelarating seismic crustal deformation before the Izmit (NW Turkey) large mainshock of 1999 August 17 and the evolution of its aftershock sequence. Geophys. J. Int. 153, 103-110.

Keilis-Borok V., Ismail –Zadeh A., Kossobokov V., Shebalin P. (2001), Nonlinear dynamics of the lithosphere and intermediate-term earthquake prediction, Tectonophysics, 338, 247-260.

Keilis-Borok V.I. (1999), What comes next in the dynamics of lithosphere and earthquake prediction, Phys. Earth Planet. Inter., 111, 179-185.

Keilis-Borok V.I. (1990), Introduction: Nonlinear systems in the problem of earthquake prediction, Phys. Earth Planet. Inter., 61, 1-7.

King, S. D., Lowman, J. P. and Gable, C. W. (2002) Episodic tectonic plate reorganizations driven by mantle convection. Earth Plan. Sci. Lett. 203, 83-91.

Kiyashchenko, D., Smirnova, N., Troyan, N., Saenger, E., Vallianatos, F. (2004) Seismic hazard precursory evolution: fractal and multifractal aspects. Phys. Chem. Earth 29, 367-378.

Kossobokov, V. G. and Shebalin, P. (2003) Earthquake prediction. In: V. I. Keilis-Borok and A. A.Soloviev (Eds) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*, Springer, Heidelberg, 141-207.

Koukouvelas, I. K., Asimakopoulos, M. and Doutsos, T. T. (1999) Fractal characteristics of active normal faults: an example of the eastern Gulf of Corinth Greece. Tectonophysics 308, 263-274.

Lana, X., Martinez, M. D., Posadas, A. M. and Canas, J. A. (2005) Fractal behaviour of the seismicity in the Southern Iberian Peninsula. Nonlin. Proces. in Geophys. 12, 353-361.

Latoussakis, J., Stavrakakis, G., Drakopoulos, J., Papanastassiou, D. and Drakatos, G. (1991) Temporal characteristics of some earthquake sequences in Greece. Tectonophysics 193, 299-310.

Lindman, M., Jonsdottir, K., Roberts, R., Lund, B. and Bodvarsson, R. (2005), Earthquakes descaled: On waiting time distributions and scaling laws. Phys. Rev. Lett. 94(10), 108501(4).

Lomnitz-Adler, J., Knopoff, L., Martinez-Mekler, G. (1992) Avalanches and epidemic models of fracturing in earthquakes. Phs. Rev.A. 45(4), 2211-2221.

Lorenz, E. N. (1991) Dimension of weather and climate attractors. Nature 353, 241-244.

Main, I. (1996) Statistical physics, seismogenesis , and seismic hazard. Rev. of Geophys. 34, 433–462.

Makropoulos, K. C. and Burton, P. W. (1984) Greek Tectonics and Seismicity. Tectonophysics 106, 275-304.

Matcharaschvili, T., Chelidze, T. and Javakhishvili, Z. (2000), Nonlinear Analysis of magnitude and interevent time interval sequences for earthquakes of the Caucasian region. Nonlin. Proces. in Geophys. 7, 9-19.

Mikhailov, A. S., and Loskutov, A. Yu. (1991) *Foundations of Synergetics II*, Springer-Verlag.

Miltenberger, P., Sornette, D., and Vanneste, C., (1993) Fault Self-organization as optimal paths selected by Critical Spatiotemporal Dynamics of earthquakes. Phys. Rev. Lett. 71(21), 3604-3607.

Molchanov, S. A., Pisarenko, V. F. and Reznikova, A. Ya. (1990) Multiscale models of failure and percolation, Phys. Earth Plan. Int. 61, 36-43.

Nakanishi, H. (1990) Cellural-automaton model of earthquakes with deterministic dynamics. Phys. Rev. A 41(12), 7086 – 7089.

Nakaya, S. and Hashimoto, T. (2002) Temporal variation of multifractal properties of seismicity in the region affected by the mainshock of the October 6 2000, Western Tottori Prefecture Japan earthquake M=7.3. Geophys. Res. Lett. 29(10), 1495.

Narkunskaya G.S. and Schnirman M.G. (1990), Hierarchical model of defect development and seismicity, Phys. Earth Planet. Inter., 61, 29-35.

Neda, Z., Bako, B. and Rees, E. (1996) The dripping faucet revisited, Chaos 6(1), 59-62.

Newman, W. I., Gabrielov, A. M., Durand, T. A., Phoenix, S. L., Turcotte, D. L. (1994) An exact renormalization model for earthquakes and material failure statics and dynamics. Physica D 77, 200-216.

Newman, W. I. and Turcotte D. L. (2002) A simple model for the earthquake cycle combining self-organized complexity with critical behaviour. Nonlin. Proces. Geophys. 9, 453-461.

Nicolis, G. (1995) *Introduction to Nonlinear Science*, Cambridge University Press, USA.

Olami, Z. and Christensen, K. (1992) Temporal correlations, universality and multifractality in a spring-block model of earthquakes. Phys. Rev. A 46(4), 1720-1723.

Olami, Z., Feder, H.J.S., and Christensen, K. (1992) Self –organized criticality in a continuous, nonconservative cellular automaton modelling earthquake. Phys. Rev. Lett. 68, 1244-1247.

Oliver, J., Sykes, L. and Isacks, B. (1969) Seismology and the new global tectonics. Tectonophysics 7(5-6), 527-541.

Ott, E. (1981) Strange attractors and chaotic motions of dynamical systems. Rev. of Modern Phys. 53 (4), 655-671.

Papadopoulos, G. A. and Dedousis, V. (1992) Fractal approach of the temporal earthquake distribution in the Hellenic Arc-trench system. PAGEOPH 139 (2), 269-276.

Papazachos, B. C. (1990) Seismicity of the Aegean and surrounding area. Tectonophysics 178, 287-308.

Papazachos, B. and Papazachos, C. (2000) Accelarated Preshock Deformation of Broad Regions in the Aegean Area. Pure Appl. Geophys. 157, 1663-1681.

Pavlos G.P., Karakatsanis L, Latoussakis J.B., Dialetis D. and Papaioannou G. (1994), Chaotic analysis of a time series composed of seismic events recorded in Japan, Intern. J. of Bifurc. And Chaos 4(1), 87-98.

Pavlos, G. P., Athanasiu M., Diamantidis D., Rigas A. G., and Sarris, E., (1999) Comments and new results about the magnetospheric chaos hypothesis, Nonlin. Proces. Geophys., 6, 99-127.

Peixoto, T.P. and Prado, C. P. C. (2006) Network of epicentres of the Olami-Feder-Christensen model of earthquakes. Phys. Rev. E 74(1), 016126(9).

Plagianakos, V. P. and Tzanaki, E. (2001) Chaotic analysis of seismic time series and short term forecasting using neural networks. IEEE, 1598-1602.

Pristley, M. B., (1988) Non-linear and non-stationary time series analysis, Academic Press.

Primakov, I. and Shnirman, M. (1999) Type of trajectory instability for a movable disk model of the lithosphere. Phys. Earth Planet. Inter. 111, 305-315.

Provenzale, A., Osborne, A. R., Kirwan, Jr. A. D., and Bergamasco, L., (1991) The study of fluid parcel trajectories in large-scale ocean flows, in Nonlinear Topics in Ocean Physics, edited by A.R. Osborne, pp. 367-402, Elsevier, Paris.

Robertson, M. C., Sammis, C. G., Sahimi, M., Martin, A. J. (1995) Fractal analysis of three-dimensional spatial distributions of earthquakes with a percolation interpretation. J. Geophys. Res. 100(B1), 609-620.

Sahimi, M., Robertson, M. C. and Sammis, C. G. (1992) Relation between the earthquake statistics and fault patterns and fractals and percolation. Physica A 191, 57-68.

Sahimi, M., Robertson, M. C. and Sammis, C. G. (1993) Fractal Distribution of earthquakes hypocenters and its relation to fault patterns and percolation. Phys. Rev. Lett. 70(14), 2186-2189.

Saichev, A. and Sornette, D. (2006) "Universal" Distribution of Interearthquake Times Explained. Phys. Rev. Lett. 97 (7), 078501(4).

Saleur, H., Sammis, C. G. and Sornette, D. (1996) Renormalization group theory of earthquakes. Nonlin. Proces. Geophys. 3, 102-109.

Shaw, R. (1984) The dripping faucet as a model chaotic system. *The science frontier express series*, Aerial, Santa Cruz.

Shebalin, P., Keilis-Borok, V., Gabrielov, A., Zaliapin, I., Turcotte, D., (2006) Short-term earthquake prediction by reverse analysis of lithosphere dynamics. Tectonophysics 413, 63-75.

Schmelling, H. and Marquart, G. (1993) Mantle flow and the evolution of the lithosphere, Phys. Earth and Plan. Int. 79, 241-267.

Schreiber, T. and Schmitz, A. (1996) Improved surrogate data for nonlinearity test. Phys. Rev. Lett. 77, 635-638.

Shcherbakov, R., Yakovlev, G., Turcotte, D. L. and Rundle, J. B. (2005) Model for the Distribution of Aftershock Interoccurence Times. Phys. Rev. Lett. 95(21), 218501(4).

Shcherbakov, R., Aalsburg, J. V., Rundle, J. B., Turcotte, D. L. (2006) Correlations in aftershock and seismicity patterns. Tectonophysics 413, 53-62.

Shnirman, M. G. and Blanter, E. M. (1999) Mixed hierarchical model of seismicity: scaling and prediction. Phys. Earth Planet. Inter. 111, 295-303.

Sornette, D. (1992) Self-organized Criticality earthquakes and Plate Tectonics. Phys. Rev Lett. 69(8), 1287.

Sornette, D. and Pisarenko, V. (2003) Fractal Plate Tectonics. Geophys. Res. Lett. 30(3), 1105.

Srivastava, H. N., Bhattacharya, S. N. and Sihna Ray, K. C. (1996), Strange attractor characteristics in Shillong plateau and adjoining regions. Geophys. Res. Lett. 23(24), 3519-3522.

Stinchcombe, R. M. (1988), Phase transitions. In: S. Lundqvist, N. March and M. P. Tosi (Ed), *Order and Chaos in Nonlinear Physical Systems*, Plenum Press, New York and London, 295.

Stiller, H., Franck, S. and Mohlmann, D. (1984) Geodynamics and state of the Earth's Interior. J. Geodynamics 1, 79-100.

Sykes, L. R., Shaw, B. E. and Scholz, C. H. (1999) Rethinking Earthquake Prediction. Pure Appl. Geophys. 155, 207-232.

Takayasu, H., (1990) Fractals in the physical sciences, Manchester University Press, Manchester and New York.

Takens, F., (1981) Detecting strange attractors in turbulence. In: D. A. Rand and L. S. Young (Eds), *Vol. 898 of Lectures Notes in Mathematics*, Spinger, Berlin,pp. 366-381.

Telesca, L., Cuomo, V., Lapenna, V., Vallianatos, F. (2000) Self-similarity properties of seismicity in the Southern Aegean area. Tectonophysics 321, 179-188.

Telesca, L., Cuomo, V., Lapenna, V., Vallianatos, F., Drakatos, G. (2001) Analysis of the temporal properties of Greek aftershock sequences. Tectonophysics 341, 163-178.

Telesca, L., Lapenna, V., Vallianatos, F. (2002) Monofractal and multifractal approaches in investigating scaling properties in temporal patterns of the 1983-2000 seismicity in the western Corinth Graben Greece. Phys. Earth and Plan. Int. 131, 63-79.

Telesca, L., Lapenna, V., Macchiato, M. (2004) Investigating linear and nonlinear behaviours in time dynamics of observational seismic sequences. Chaos, Solitons and Fractals 20, 195-203.

Temam, R. (1988) *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*. Springer-Verlag, New York.

Theiler, J. (1991) Some comments on the correlations dimensions of $1/f^{a}$ noise. Phys. Lett. A 155, 480-493.

Theiler, J., Galdikian, B., Longtin, A., Eubank, S. and Farmer, J. D. (1992a) Using surrogate data to detect nonlinearity in time series. In: M. Casdagli and S. Eubank (Eds) *Nonlinear Modeling and Forecasting, vol. XII of SFI studies in the Sciences of Complexity*. Addison-Wesley, Reading, Mass, p. 163-188.

Theiler, J., Eubank, S., Longtin, A., Galdikian, B. and Farmer, J. D. (1992b) Testing for nonlinearity in time series: the method of surrogate data. Physica D 58, 77-94.

Tiwari, R. K., Lakshmi, S. Sri and Rao, K. N. (2004), Characterization of Earthquake Dynamics in Northeastern India Regions: A Modern Forecasting Approach. Pure Appl. Geophys. 161, 865-880.

Tiwari, R. K. and Lakshmi, S. Sri (2005) Some common and contrasting features of earthquake dynamics in major tectonic zones of Himalayas using nonlinear forecasting approach. Current Science 88(4), 640-647.

Tsonis, A. A., (1992) *Chaos: from theory to applications*. Plenum, New York.

Turcotte, D. L. (1986) A fractal model for crustal deformation. Tectonophysics 132, 261-269.

Turcotte, D. L. (1999) Seismicity and self-organized criticality. Phys. Earth Planet. Inter. 111, 275-293.

Turcotte, D. L., Malamud, B. D., Morein, G., Newman, W. I. (1999) An inverse cascade model for self-organized critical behaviour. Physica A 268, 629-643.

Turcotte, D. L. (2001) Self-organized criticality: Does it have anything to do with criticality and is it useful? Nonlin. Proc. Geophys. 8, 193-196.

Urquizu, M. and Correig, A.M. (1998) Analysis of seismic dynamical systems. J.of Seismology 2, 159-171.

Utsu, T. (1970) Aftershocks and earthquake statistics (II). Geophysics 3, 197-226.

Vasconcelos G.L., Vieira M. S. and Nagel S.R. (1992), Phase transitions in a spring-block model of earthquakes, Physica A, 191, 69-74.

Vieira, M. S. (1992) Self-organized criticality in a deterministic mechanical model. Phys. Rev. A 46(10), 6288-6293.

Vieira, M. S., Vasconcelos, G. L. and Nagel, S. R. (1993), Dynamics of spring-block models: Tuning to criticality, Phys. Rev. E 47(4), 2221-2224.

Vieira, M. S. and Herrmann, H. J. (1994) Self-similarity of friction laws. Phys. Rev. E 49(5), 4534 - 4541.

Vieira, M. S. and Lichtenberg, A. J. (1996) Presence of chaos in a self-organized critical system. Phys. Rev. E 53(2), 1441 – 1445.

Vieira, M. S. (1999) Chaos and synchronized Chaos in an Earthquake Model. Phys. Rev. Lett. 82(1), 201-204.

Vieira, M. S. (2000) Simple deterministic self-organized critical system. Phys. Rev. E 61(6), 6056-6059.

Vieira, M. S. (2002) Breakdown of self-organized criticality in sandpiles. Phys. Rev. E 66, 051306(5).

Vieira, M. S. (2004) Are avalanches in sandpiles a chaotic phenomenon. Physica A 340, 559-565.

Weatherley, D. (2006) Recurrence Interval Statistics of Cellural Automaton Seismicity Models. Pure Appl. Geophys. 163, 1993-1947.

Wolf, A., Swift, J. B., Swinney, H. L., Vastano, J. (1985) Determining Lyapunov exponents from a time series. Physica D 16, 285-317.

Woodard, R., Newman, D., E., Sanchez, R., Benjamin, A., C. (2007) Persistent dynamic correlations in self-organized critical systems away from their critical point. Physica A 373, 215-230.

Xu, Y. and Burton, P. W. (1995) Temporal Scaling Regions and Capacity Dimensions for Microearthquake Swarms in Greece. In: Novak M. M. (Ed), *Fractal Reviews in the Natural and Applied Science.* Chapman and Hall, London, pp. 292-300.

Xu, Y. and Burton, P. W. (1999) Spatial Fractal Evolutions and Hierarchies for Microearthquakes in Central Greece. Pure Appl. Geophys. 154, 73-99.

Xu, Y. and Burton, P. W. (2006) Time varying seismicity in Greece: Hurst's analysis and Monte Carlo simulation applied to a new earthquake catalogue for Greece. Tectonophysics 423, 125-136.

Yang, X., Du, S. and Ma, J. (2004) Do Earthquakes exhibit Self-Organized criticality. Phys. Rev. Lett. 92(22), 228501(4).

15 Analysis of Nonlinear Biophysical Time Series in Aquatic Environments: Scaling Properties and Empirical Mode Decomposition

François G. Schmitt[1], Yongxiang Huang[1,2], Zhiming Lu[2], Sylvie Zongo Brizard[1], Juan Carlos Molinero[3] and Yulu Liu[2]

[1] ELICO laboratory, FRE CNRS n°2816, University of Science and Technologies of Lille – Lille 1, Francois.schmitt@univ-lille1.fr
[2] SIAMM, University of Shanghai, Shanghai, R. P. China, lanlankai@163.com
[3] UMR CARRTEL, INRA, Thonon-les-Bains, France, molinero@thonon.inra.fr

Abstract. Aquatic environmental time series often display large fluctuations at many time scales, possessing stochastic properties, as well as deterministic forcing coming from seasonal or annual meteorological and climatic cycles. In this work we are interested in the characterization of these properties, using different statistical tools, borrowed from the field of turbulence, or of nonlinear time series analysis. We first present the analysis of a long (30 years) time series of daily river flow data, recorded in the Seine River (France). We consider the scale dependence and scale invariance of river flow data, using structure function analysis; we also apply a decomposition method called Empirical Mode Decomposition (EMD). We then consider the statistical properties, and the nonlinear dynamics behaviour of a long-term copepod (small crustaceans) time series sampled every week in the Meditarranean sea from 1967 to 1992. We first consider its high variability and characterize its properties, including extreme evens obeying power law tail pdf. We then consider their scale dependence, using Fourier power spectra together with an EMD approach.

1 Introduction

A general feature of environmental and geophysical fields is their high variability at many different time and space scales. Here we consider long-term time series, recorded at daily and weekly sampling intervals over several decades. The time scales dealt with such time series include the annual cycle, which often provides a strong forcing apparent as a pike in the power spectra. Such data therefore present large fluctuations at many time scales possessing stochastic properties, together with a deterministic forcing coming from the annual cycle. We apply here two methods to describe the scale dependence of the fluctuations. We consider the structure functions approach, borrowed from the field of turbulence (Frisch 1995) and the Empirical Mode Decomposition method, a recently introduced time series analysis algorithm (Huang, Shen, Long, Wu, Shih, Zheng, Tung and Liu 1998; Huang, Shen and Long 1999). We first present in section 2 below, the analysis of a long (30 years) time series of daily river flow data, recorded in the Seine River (France). We then

consider in section 3, a long-term copepod time series sampled every week in the Meditarranean sea from 1967 to 1992. In each case, we first present the database, followed by statistical analysis, scaling properties, and EMD decomposition.

2 Analysis of the Seine River Flow Data

A precise understanding of river runoff processes is important for many applications including floods and droughts forecasting and estuarine ecosystem studies. It is recognized for a long time that river flow fluctuations possess scaling properties and long-range statistical dependencies (Hurst 1951). More recently these fluctuations have been studied in various scaling frameworks, including Farima models (Montanari, Rosso, Taqqu 1997), universal multifractals (Tessier, Lovejoy, Hubert, Schertzer, Pecknold 1996; Pandey, Lovejoy, Schertzer 1998) or multifractal detrended fluctuation analysis (MFDA) (Kantelhardt, Rybski, Zschiegner, Braun, Koscielny-Bunde, Livina, Havlin, Bunde 2003; Koscielny-Bunde, Kantelhardt, Braun, Bunde, Havlin 2006). These studies considered daily river flow data, and characterized the scaling ranges observed, superposed to the annual cycle. Here we perform the analysis of another daily river flow time series (the Seine river, France), with new methods (structure functions and EMD) to characterize stochastic fluctuations and scaling regimes.

2.1 Presentation of the data

We analyse here Seine river flow data provided by the Service de Navigation de la Seine, France (SNS). This corresponds to daily flow data Q (in m^3/s), recorded from 1 January 1941 to 28 May 2006. There are 23 008 data values, with some missing values due to some interruptions, for maintenance or because of the failure of measuring devices. The maximum interruption length is an interruption of 801 days during the Second World War. There are also two interruptions of one month duration. The data are shown in Fig. 15.1, showing the large fluctuations at all scales. The mean discharge is $q_0 =< Q >= 436.4$ m^3/s, with some large values, the largest being $q_{\max} = 2280$ m^3/s. A zoom on a two-year duration (1980-1981) is shown in Fig. 15.2. This illustrates the complex and stochastic behaviour, with a visible strong annual cycle and small-scale stochastic fluctuations. The standard deviation of the data is large, close to the mean value (330 m^3/s), indicating a dissymmetric probability density function.

2.2 Scaling analysis and structure functions

We characterize here the dynamics of the time series. To estimate the Fourier power spectrum, the classical FFT algorithm needs a regular sampling. The original time

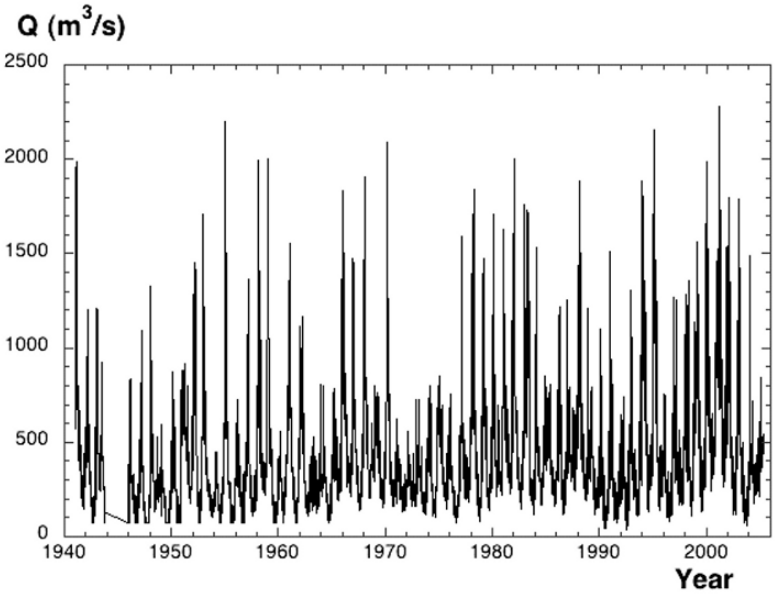


Fig. 15.1. The river flow time series, illustrating the huge fluctuations, with frequent and large floods of more than 1000 m^3/s

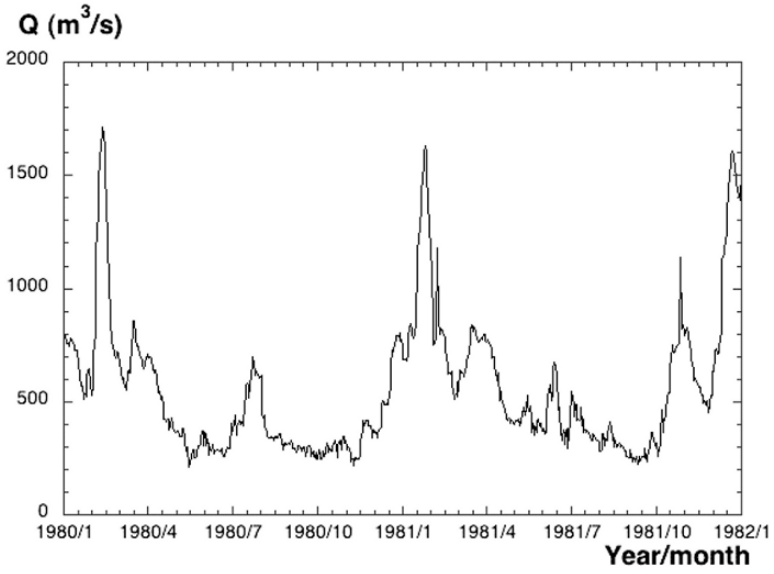


Fig. 15.2. The river flow time series for a duration of two years (1980 and 1981), emphasizing the annual cycle and small-scale stochastic fluctuations

series with missing values is thus filled with zero values: this provides a new time series with a regular sampling, without changing the power spectral estimation (the Fourier transform is not modified by introducing zeroes). The resulting power spectrum is shown in Fig. 15.3, in log-log plot, to emphasize power-law regimes. To ensure a smooth power spectrum, it has been bin-averaged, smoothing large frequency fluctuations. There is a rough general decreasing trend, with a strong annual cycle. We have emphasized two scaling trends of the form $E(f) \approx f^{-\beta}$, with a slope of $\beta \approx 1.5$, for scales between 1 year and 1 month (frequencies between 1 and 15 year^{-1}), and for scales between 1 and 15 days (frequencies larger than 25 year^{-1}).

The power spectrum is a second order moment; it corresponds to estimate the frequency – and hence time scale – dependence of medium fluctuations. In the following we will consider the whole range of fluctuations (small, medium and large) using statistical moments of river discharge fluctuations, at various scales: for this, we estimate structure functions, an analysis technique borrowed from the field of turbulence (see Frisch 1995), which has already been applied to other time series such as climate data (Schmitt, Lovejoy and Schertzer 1995). We consider fluctuations of discharge data, written as $\Delta Q_\tau = |Q(t+\tau) - Q(t)|$. For scaling processes,

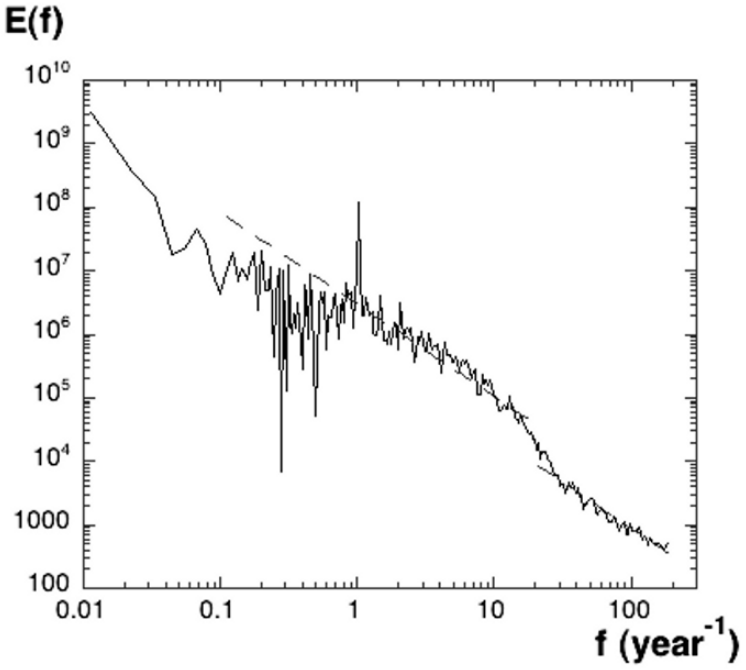


Fig. 15.3. The Fourier power spectrum of the data (bin-averaged), in log-log plot, emphasizing the roughly scaling regimes (dotted lines, both of slope -1.5), and a neat and strong pike associated to the annual cycle.

intermittency of the structure functions can be written (Frisch 1995; Schmitt et al. 1995):

$$< (\Delta Q_\tau)^p >=< (\Delta Q_T)^p > \left(\frac{\tau}{T}\right)^{\zeta(p)} \quad (1)$$

where T is the external scale of the scaling regime and $\zeta(p)$ is a scaling moment function characterizing the scaling properties of the discharge fluctuations: for monofractal processes, it is linear, and for multifractal processes, it is nonlinear and concave; for example for a lognormal multifractal process, this function is quadratic. The first moment gives the scaling exponent $H = \zeta(1)$ corresponding to the scaling of the average fluctuations. We note $m(p,\tau) =< (\Delta Q_\tau)^p >$, and first consider $g(\tau) \equiv m(2,\tau)$. Its scale dependence is represented in log-log plot in Fig. 15.4. It has a nontrivial scale behaviour: there is a roughly scaling behaviour for small time differences; for time scales close to one year, there is clear influence of the annual cycle. The dotted line indicates a limited scaling range, with a slope of 0.68. Relation (1) expresses an exact power-law relation for the moments of the fluctuations. A weaker scaling relation has been proposed in the turbulence literature: Extended Self-Similarity (ESS) (Benzi et al 1993). This corresponds to express relative scaling properties: the moments $m(p,\tau)$ are simply related to $m(2,\tau)$:

$$m(p,\tau) = F(T,p) g(\tau)^{\Psi(p)} \quad (2)$$

where $F(T,p)$ is a constant that does not depend on τ and $\Psi(p)$ is a scaling moment function. When (1) is valid, we have $\Psi(p) = \zeta(p)/\zeta(2)$. A relation such as (2) captures the intermittent fluctuations of the series through the function $\Psi(p)$, and its scale dependence for all order of moments (and hence all fluctuation intensities) is given by the τ-dependence of $g(\tau) \equiv m(2,\tau)$. Figure 15.5 represents $m(p,\tau)$ versus $g(\tau)$ in log-log plot, for p=1,3,4. The nice straight lines which are obtained confirm the validity of the relation (2) for the smaller scales. This has been done for many different values of p (even non-integer values) from 0 to 5. The resulting function $\Psi(p)$ is shown in Fig. 15.6. It is compared to a straight line of equation $q/2$: the nonlinear and concave curve which is obtained confirms the multifractal properties of river flow fluctuations, as obtained previously by other methods applied to daily river flow data (Tessier et al. 1996; Pandey et al. 1998; Kantelhardt et al. 2003; Koscielny-Bunde et al. 2006).

Within this structure function approach, we may say that the knowledge of the function $g(\tau) \equiv m(2,\tau)$ and $\Psi(p)$ provide a characterization of all fluctuations at all scales below one year. For larger scales, we consider below the deterministic trend using the decomposition method proposed by the EMD algorithm.

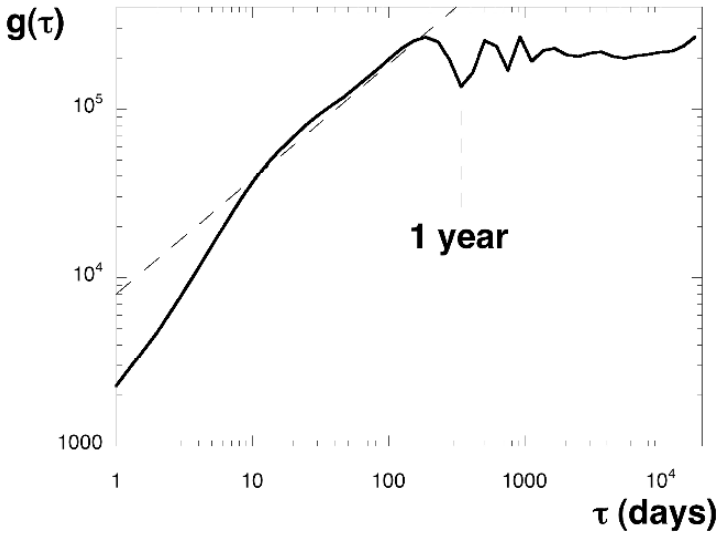


Fig. 15.4. The function $g(\tau) \equiv m(2,\tau)$. There is a roughly scaling behaviour for small time differences; the straight line of slope 0.68 indicates the range of validity of the scaling behaviour. For time scales close to one year, there is clear influence of the annual cycle.

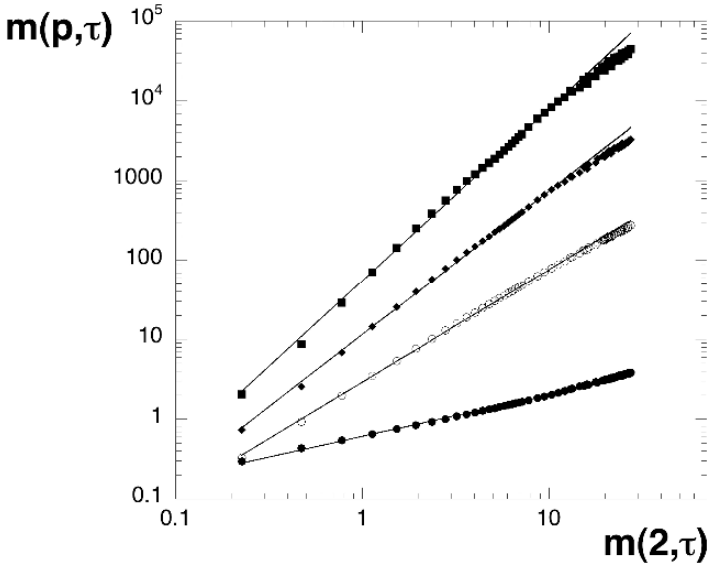


Fig. 15.5. $m(p,\tau)$ versus $g(\tau)$ in log-log plot, for p=1,3,4 (from bottom to top). The nice straight lines obtained confirm the validity of Equation (2) for the smaller scales.

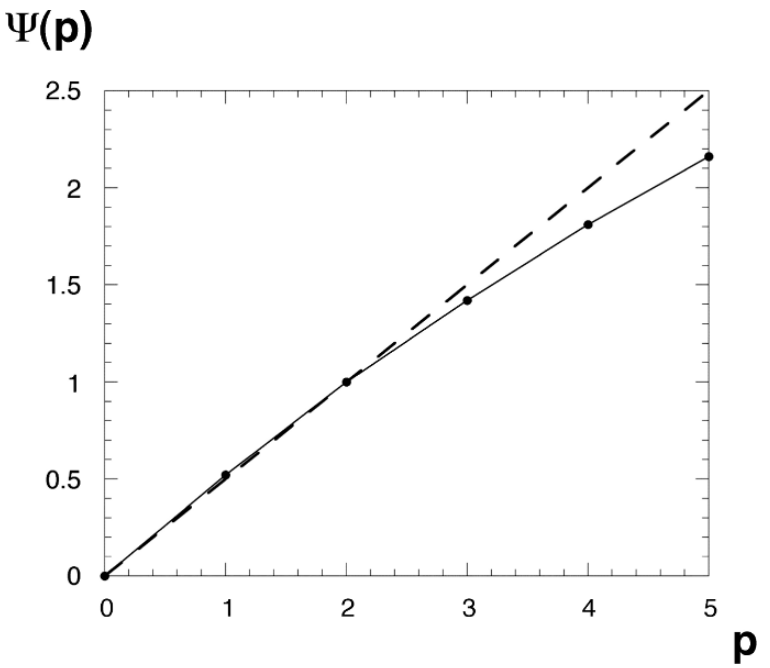


Fig. 15.6. The function $\Psi(p)$ estimated from the data (dots and continuous line). A straight line of equation $p/2$ is shown for comparison (dotted line).

2.3 Empirical mode decomposition of the Seine river flow data

Empirical Mode Decomposition (EMD) is a recently developed method (Huang, Shen, Long, Wu, Shih, Zheng, Tung and Liu 1998; Huang, Shen and Long 1999) that can be applied to study the nonlinear and non-stationary properties of a time series. The starting point of EMD is to locally estimate a signal as a sum of a local trend and a local detail: the local trend is a low frequency part, and the local detail a high frequency. When this is done for all the oscillations composing a signal, the high frequency part is called an Intrinsic Mode Function (IMF) and the low frequency part is called the residual. The procedure is then applied again to the residual, considered as a new times series, extracting a new IMF and a new residual. At the end of the decomposition process, the EMD method expresses a time series as the sum of a finite number of IMFs and a final residual (Huang et al. 1998; Flandrin, Rilling and Goncalves 2004). The procedure is precisely described below.

An IMF is a function that satisfies two conditions: (i) the difference between the number of local extrema and the number of zero-crossings must be zero or one; (ii) the running mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. The procedure to decompose a signal into IMFs is the following (Huang et al. 1998; Huang et al. 1999):

1) The local extrema of the signal $x(t)$ are identified;
2) The local maxima are connected together forming an upper envelope $e_{\max}(t)$, which is obtained by a cubic spline interpolation. The same is done for local minima, providing a lower envelope $e_{\min}(t)$;
3) The mean is defined as $m_1(t) = \left(e_{\min}(t) + e_{\max}(t)\right)/2$;
4) The mean is subtracted from the signal, providing the local detail $h_1(t) = x(t) - m_1(t)$;
5) The component $h_1(t)$ is then examined to check if it satisfies the conditions to be an IMF. If yes, it is considered as the first IMF and denoted $C_1 = h_1$. It is subtracted from the original signal and the first residual, $r_1(t) = x(t) - C_1(t)$ is taken as the new series in step 1. If $h_1(t)$ is not an IMF, a procedure called "sifting process" is applied as many times as needed to obtain an IMF. In the sifting process, $h_1(t)$ is considered as the new data; the local extrema are estimated, lower and upper envelopes are formed and their mean is denoted $m_{11}(t)$. This mean is subtracted from $h_1(t)$, providing $h_{11}(t) = h_1(t) - m_{11}(t)$. Then it is checked if $h_{11}(t)$ is an IMF. If not, the sifting process is repeated, until the component $h_{1k}(t)$ satisfies the IMF conditions. Then the first IMF is $C_1 = h_{1k}$ and the residual $r_1(t) = x(t) - C_1(t)$ is taken as the new series in step 1.

By construction, the number of extrema decreases when going from one residual to the next; the above algorithm ends when the residual has only one extrema, or is constant, and in this case no more IMF can be extracted. The complete decomposition is then achieved in a finite number of steps, of the order $n \approx \log_2 N$, for N data points (Flandrin and Goncalves 2005).

The signal $x(t)$ is finally written as

$$x(t) = \sum_{i=1}^{n} C_i(t) + r_n(t) \tag{3}$$

The IMFs are orthogonal, or almost orthogonal functions, meaning that they are mutually uncorrelated. This method does not require stationarity of the data and is especially suitable for nonstationary and nonlinear time series analysis (Huang et al. 1998; Huang et al. 1999). Each modes are localized in frequency space (Flandrin and Goncalves 2004; Wu and Huang 2004). EMD is a time-frequency analysis (Flandrin et al. 2004) since it can represent the original signal in a energy-frequency-time form at local level, using a complementary method called Hilbert-Huang spectrum (Huang et al. 1998). This decomposition can be used to express the original time series as the sum of a trend (sum of modes from p to n) and small-scale fluctuations (sum of modes from 1 to p-1), where p is an index whose value depends on the trend decomposition which is desired.

Due to the simplicity of its algorithm, the EMD method has met a large success, and many methodological papers (e.g. Wu and Huang 2004; Delechelle et al. 2005;

Flandrin et al. 2004; Flandrin and Goncalves 2005) have appeared recently. This technique has already been applied to several fields, including acoustics (Loutridis 2005), biomedical applications (Ponomarenko et al. 2005), climate (Salisbury and Wimbush 2002; Coughlin and Tung 2004) and nonlinear waves in oceanography (Hwang et al. 2003; Veltcheva and Guedes Soares 2004). As far as we know, it has never been applied to aquatic ecosystem time series.

Here we present some first results applying EMD to the daily Seine river flow data. Figure 15.7 shows the successive IMFs obtained from the original time series, for modes from 1 to 13 with the last residual. As analyzed by Flandrin and Goncalves (2005) and Wu and Huang (2004), each mode is localized in frequency space. The mean time scale associated to each mode (from 1 to 13) is shown in Table 15.1; it is estimated by considering the (energy weighted) mean frequency associated to each mode. We can see that the modes are regularly spaced, with the 9th being close to the annual cycle, and the 7th to the seasonal time scale. A phenomenon of mode mixing (Huang et al. 1998; Huang et al. 1999) is observed in the 8th mode: part of it should belong to 9th mode. After decomposition, we have therefore separated the original time series into two terms: small-scale fluctuations corresponding to the sum from

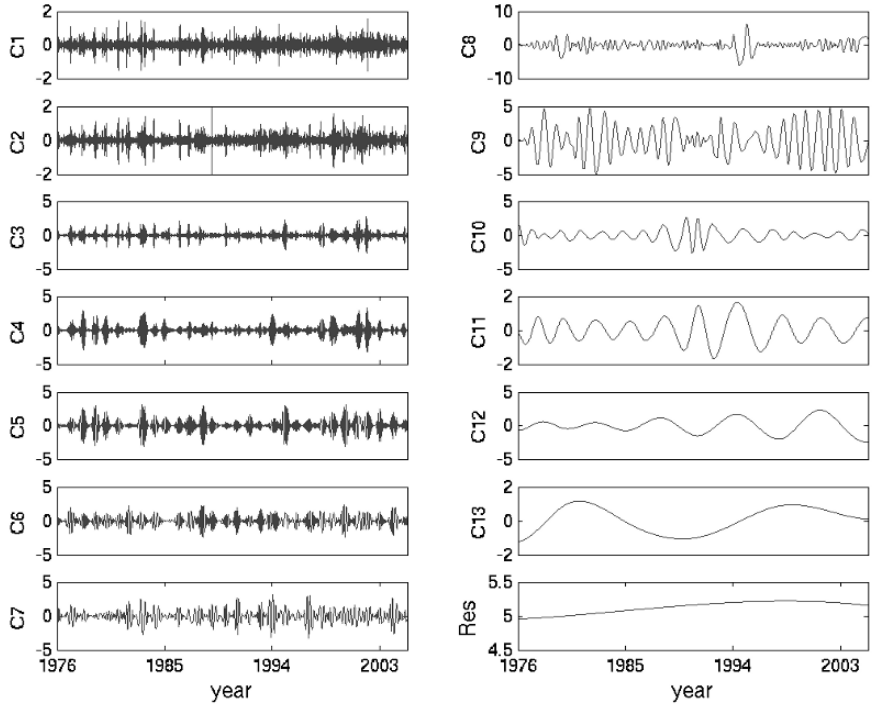


Fig. 15.7. The first 7 IMFs components of Seine River from EMD: the time scale increase with the order of the modes and the small scales are strongly fluctuating. Mode mixing (Huang et al. 1998; Huang et al. 1999) is observed in the 8th mode.

mode 1 to 7; and the trend, corresponding to all the other modes including the 8th mode (large scales and the residual). A portion of each of these two series is shown in Fig. 15.8, superposed with the original data. It can be seen that the trend nicely captures the large-scale behaviour, whereas the small-scale data may be analyzed and characterized using stochastic methods, as done above using the structure functions approach. Their corresponding underlying dynamical mechanism may be different: the methods applied here may provide a path to establish models for small scale or large-scale fluctuations. This will be studied further in future studies.

Table 15.1. Mean period (in days) of each mode: mode 9 shows clearly a strong annual cycle. Due to mode mixing, it is better to include the 8th mode into the trend.

Mode	1	2	3	4	5	6	7	8	9	10	11	12	13
MP	3	7	17	29	49	69	101	296	377	534	1127	2389	4338

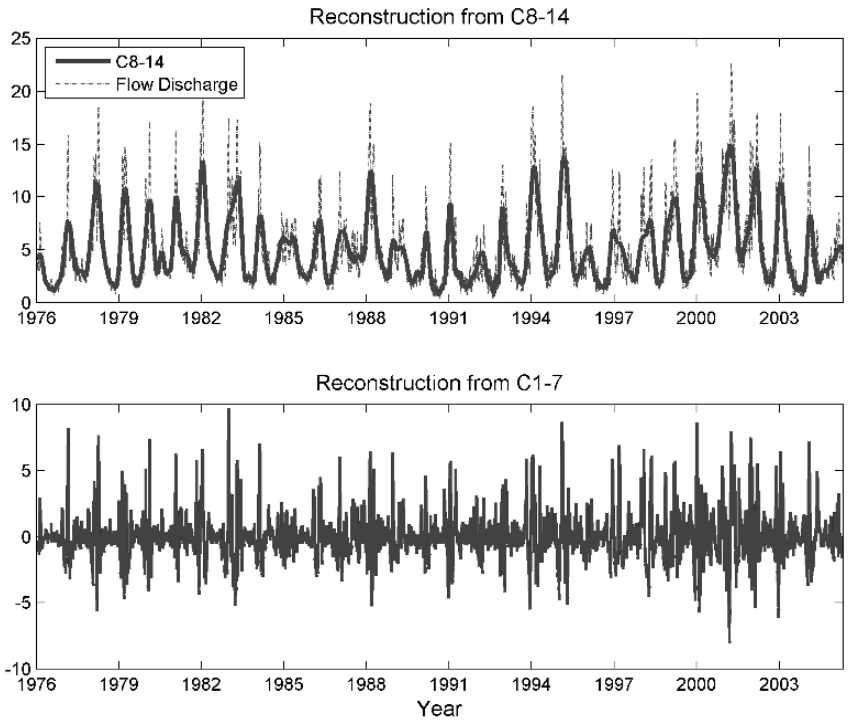


Fig. 15.8. Reconstruction from the IMF components: (a) original data (dotted line) and the sum of C8-14 (solid line), which means periods close and larger than annual cycle; (b) the sum of C1-7 illustrating small scales turbulent-like fluctuations.

3 Analysis of a Copepod Long-term Time Series

Population dynamics of natural populations is influenced by the environment, and possesses a complex dynamics due to internal processes (i.e. fecundity, mortality) and external forcing factors (i.e. ecological interactions, environmental variability). This provides complex couplings, at the origin also of high variability at many different time and space scales for animal population evolution. This was noted very early, as shows the following citation from Darwin (1859): "What checks the natural tendency of each species to increase in number is most obscure." The ecologist Charles Elton was one of the first to remark that despite huge variability, some populations display a periodic variability (Elton 1924). Following this, Pat Moran proposed a density-dependent forcing model of the environment on the population dynamics (Moran 1953). Other models have been proposed, many of them being deterministic at the macroscopic level: exponential mortality or growth, predator-prey models, etc. (see Frontier et al. 2004), but huge stochastic fluctuations visible on population time series are not always compatible with such deterministic dynamical models; indeed these models are justified for closed systems, whereas natural populations in turbulent environments would need an open system description with many degrees of freedom. Here we use stochastic approaches for a population dynamics time series, and characterize some of its properties. As before, we first consider scaling methods, before testing the EMD.

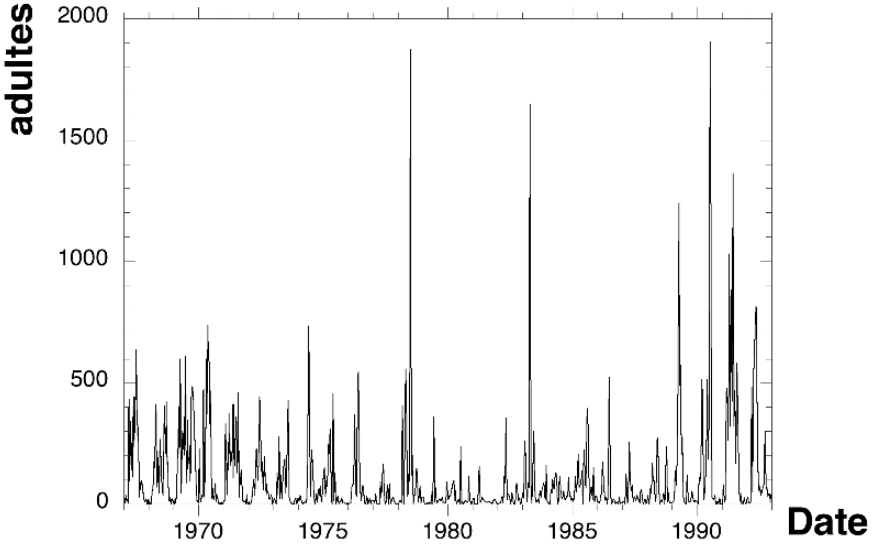


Fig. 15.9. Time series of the weekly copepod abundance between 1967 and 1992: huge fluctuations are clearly visible.

3.1 Presentation of the data

It is not always possible to measure the total abundance of a population. When one deals with aquatic ecology data, the temporal variability of the population can be sampled with regular measurements, at a fixed position, using the same measuring procedure. This does not measure the total population, and is therefore not optimal, but it is hoped that this under-sampling is representative of the dynamics of the population. This undersampled information performed during several decades at the same position and using the same procedure, is in fact precious, and the objective of the present data analysis approach, is to extract as much as possible information on the dynamics and statistics of the series.

In ocean ecosystems planktonic copepods constitute the majority of the mesozooplankton biomass. Their secondary production supports most of the food webs in the open sea directly affecting pelagic fish populations and the biological pump of carbon into the deep ocean (Ohman and Hirche 2001). Moreover, these organisms have a typical size of the order of milimeters, a fast life cycle (about one month) and are usually considered to be deeply adapted to their turbulent environment, while being also sensitive to variability in the environmental properties, such as temperature, turbidity, turbulence intensity.

We analyze here a long-term database of *Centropages typicus* abundance, sampled every week from 1967 to 1992, focusing on the nonlinear dynamics of the population and its extremes. In a previous study (Schmitt et al 2007) we have focused on the male-female dynamics of this population. The Villefranche Bay is situated in the

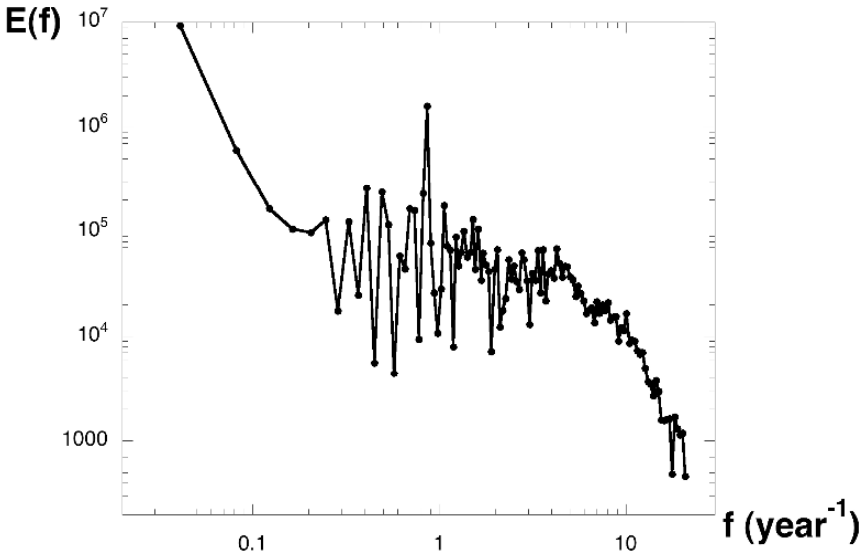


Fig. 15.10. Power spectrum of the copepod time series. A pike corresponding to the annual cycle is clearly visible.

Ligurian Sea, Northwestern Mediterranea. In the mouth of the bay, a monitoring of plankton variability has been carried out by the Oceanographic Observatory of Villefranche, France (sampling station Pt B located at 43 deg 41' N; 7 deg 19' E) since January 1967. During daylight, plankton sampling was carried out by vertical hauls using a Juday-Bogorov net (mouth aperture 0.5 m, filtering length of 1.80 m and mesh size of 330 μm) from the bottom to the surface (0-75 m). Counts of *C. typicus* adults (males and females) were made under a dissecting microscope.

Data we report correspond to the period of January 1967 to December 1992. There are 1353 data points, with a minimum value of 0 and a maximum of 1907 individuals/m^3. The mean is 108, the median is 35, indicating a strong dissymmetry in the pdf, with a large standard deviation (185) and skewness (4.02). The resulting time series for total abundance is shown in Fig. 15.9: huge fluctuations are clearly visible, indicating the intermittency properties of the time series.

3.2 Power spectrum of the dynamics and extremes

We then estimate the power spectrum of the copepod time series, represented in Fig. 15.10, in log-log plot, to emphasize possible power-law regimes. As for the river flow data, to ensure a smooth power spectrum, it has been bin-averaged, smoothing large frequency fluctuations. A strong annual cycle, and no general scaling range are visible. For larger frequencies, and smaller times scales, there is a rather fast decrease of the energy density. We therefore do not perform here a scaling analysis of

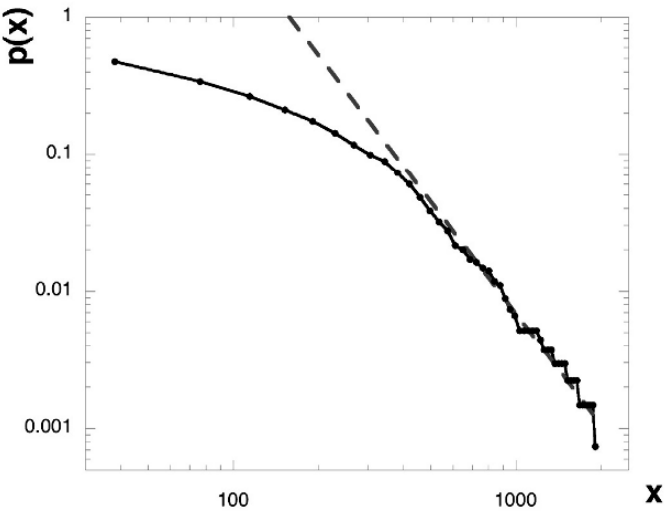


Fig. 15.11. The pdf of copepod abundance, in log-log plot. This display emphasizes the heavy tails of experimental data: it shows that the extremes are much more frequent than for a Gaussian variable. The dotted line corresponds to a hyperbolic pdf with a power-law slope of $-\mu-1$, with $\mu=2.5$.

the fluctuations. We rather focus on the probability density function (pdf) and on the extremes on the series.

The probability density function of abundance time series is displayed in Fig. 15.11. Since we focus here on extremes, this pdf is represented in log-log plot, emphasizing the tail behaviour. The tail is of the form:

$$p(x) = \frac{C_0}{x^{\mu+1}}; \quad x >> 1 \tag{3}$$

Were C_0 is a constant, x a threshold value, and μ>0 a parameter characterizing the power-law tail. Such distributions have been given many names in the literature, depending on the fields, called "power-laws", "hyperbolic distributions", or "Pareto". In the mathematical theory of extreme events (Leadbetter 1983), these laws correspond to "type II" and are sometimes called "Frechet laws". Such tail behaviour is much heavier than Gaussian or even lognormal distributions, indicating that extreme events are much more frequent. Power-law distributions have been found in many applications in the fields of natural or human sciences: for general reviews, see (Mitzenmacher 2003; Newman 2005). Hereafter we denonote such tail behaviour as power-law or hyperbolic.

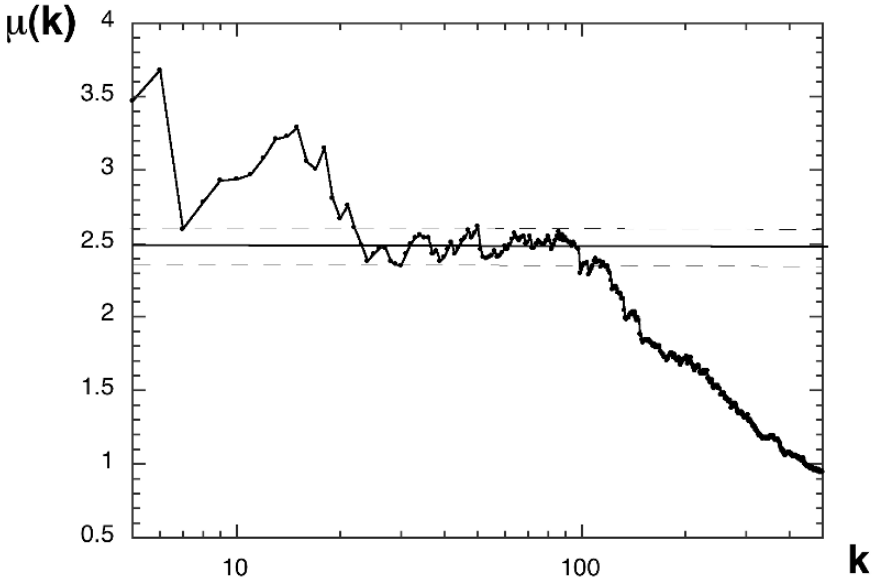


Fig. 15.12. Hill estimator applied to the total adult population of *C. typicus*: for small values of k, the estimator fluctuates due to poor amount of data; for a medium range of k values, the estimator is very stable, providing a value of $\mu = 2.5 \pm 0.05$. For larger values of k, the estimator reaches values that do not belong to the hyperbolic tail, and hence it decreases.

The value of the parameter μ is important to characterize such distribution. It is a critical order of divergence of moments. Its value can be provided using an estimator (Hill 1975; Caers 1998; Peng 1998). A classical estimator is Hill's estimator (Hill 1975), defined introducing an index $\mu(k)$, for an integer $k \geq 1$:

$$\mu(k) = \frac{k}{\sum_{i=1}^{k} \log \frac{X_{[i]}}{X_{[k+1]}}} \tag{4}$$

where the original time series has been ordered from the largest value to the smallest, the result being denoted $X_{[i]}$:

$$X_{[1]} \geq X_{[2]} \geq X_{[3]} \ldots \geq X_{[k]} \geq \ldots \geq X_{[n]} \tag{5}$$

This estimator (Equation (4)) corresponds to compute an estimation of the mean slope of the last *k* points of the pdf. In case of pure power-law pdf, this estimator provides a constant value $\mu(k) = \mu$, whereas for experimental data this estimator is expected to be roughly constant for a range of values of *k*, providing a quantitative estimation of the parameter μ. When *k* is too small, a scatter is expected for $\mu(k)$; when the hyperbolic law is valid only for extreme events, the estimator will fall down for large values of *k*. This estimator is applied on the total population in Fig. 15.12: a plateau is reached for a range of values of *k*. For these values, we have the following estimation: $\mu = 2.5 \pm 0.05$. This value is quite small; it indicates that the variance of the population is defined, but theoretically, its skewness diverges. In practice, this means that the skewness value depends on the number of realizations taken into account: the larger the number of realizations, the larger the numerical estimation of the skewness (see Schmitt and Nicolis (2002) for a discussion in the context of high frequency precipitation data analysis). Such fit of the pdf is also interesting for the purpose of modelling of population dynamics. Further work is necessary to check if this property is a universal property of copepod population abundance, or even of marine organisms population abundance.

3.3 EMD analysis

We have also applied EMD analysis on the copepod time series. This is expected to be a useful approach for the analysis of such data, possessing small-scale stochastic fluctuations superposed to a strong annual cycle. Figure 15.13 shows the successive IMFs obtained from the original time series. As done above for the river flow data, the mean periods corresponding to each mode (from 1 to 10) are shown in Table 15.2, estimated from the energy weighted Fourier power spectrum of each mode. We see also that the modes are regularly spaced, with the 5th being very close to the annual cycle. Due to mode mixing, the scale of the 4th mode is also close to the 1 year time scale; therefore, in the trend analysis we have considered small-scale fluctuations as corresponding to modes from 1 to 3. We have thus separated the original time series into two terms: small-scale fluctuations (the sum from mode 1 to 3) and the trend,

corresponding to all the other modes (large scales and the residual). A portion of each of these two series is shown in Fig. 15.14, superposed with the original data. It can be seen here again that the trend nicely captures the large-scale behaviour. This can be a useful approach for modeling population dynamics at scales larger than one year, and for studying the small-scale stochastic dynamics without being influenced by the strong large-scale cycle. This type of analysis will be done in future studies.

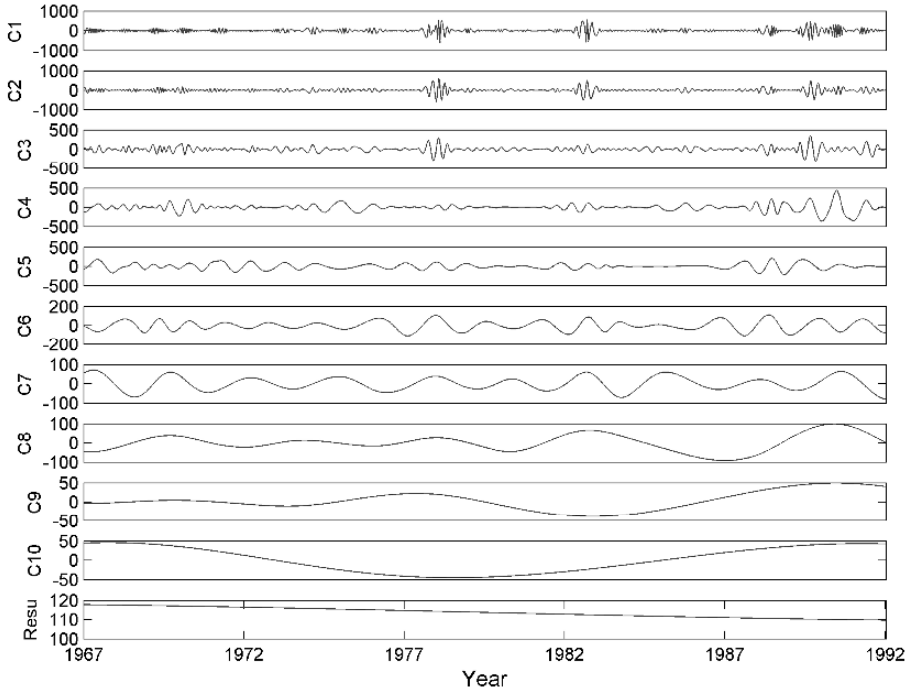


Fig. 15.13. IMF components of the copepod population: the original time series is decomposed into 10 modes having different mean time scales.

Table 15.2. Mean period (in weeks) of each mode (excluding the residual): mode 5 corresponds to the annual cycle.

Mode	1	2	3	4	5	6	7	8	9	10
MP	6	10	17	39	52	79	126	314	641	1302

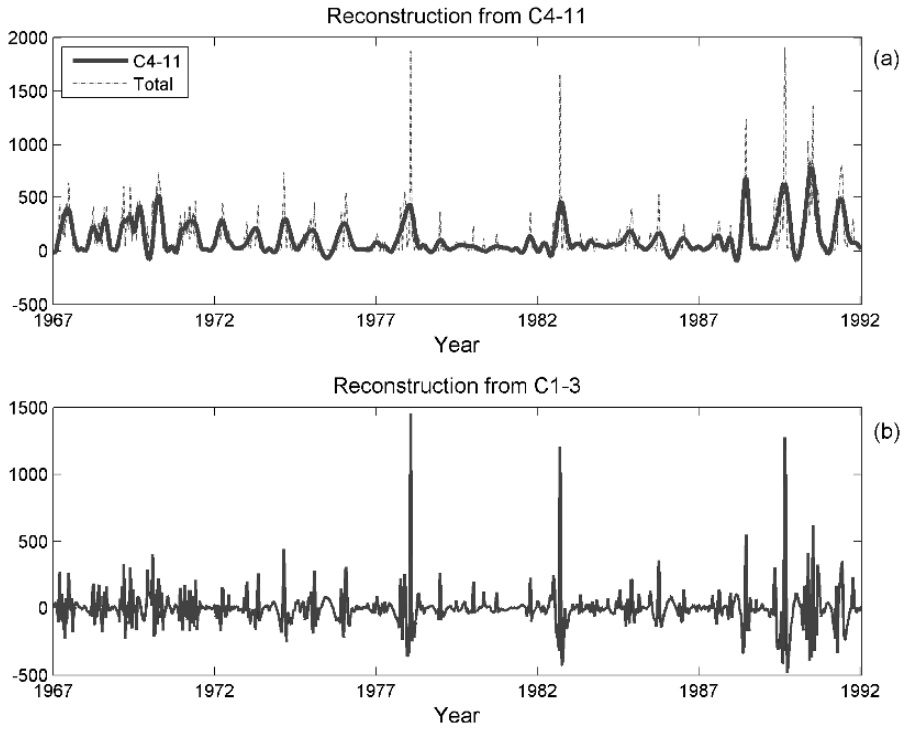


Fig. 15.14. Reconstruction from the IMF components: (a) original data (dotted line) and the sum of C4-11 (solid line), which corresponds to periods close and larger than the annual cycle; (b) the sum of modes C1-3 illustrating the small scales turbulent-like fluctuations.

4 Conclusion

We have considered here environmental time series possessing large fluctuations at many time scales, characterized by stochastic fluctuations superposed to strong annual forcing: daily Seine river flow data recorded during several decades, and a long-term biological time series recorded weekly during several decades. We have estimated the Fourier power spectra of both series; we have also applied the empirical mode decomposition method, and showed that it nicely captures in both cases the annual cycle. This method could then be used to separate the series into a large scale trend associated to the annual deterministic forcing, and small scale fluctuations. We also showed that small-scale river flow data could be analyzed using the structure functions approach, an analysis technique inspired by the field of turbulence: using the extended self-similarity framework, we could show that small-scale fluctuations have multifractal statistics, with relative multiscaling properties. This provides a multi-scale characterization of river flow intermittency. The intermittency of the

copepod time series could only be characterized through its probability density function, revealing a probability tail that follows an algebraic decay.

The investigation of extreme events has numerous implications in ecology since they shape the physiology, ecology and evolution of organisms. Some relevant examples in ecology are provided by harmful algal blooms, microphytobenthos patches when power pdfs have been found for spatial fluctuations (Seuront and Spilmont 2002), large-fire disruptions, extinctions of populations (Katz 2005). There has been recently an increasing awareness of the role of extreme events in aquatic ecology. The approach we used in this work is a contribution towards a better understanding of the dynamics of extreme events. A further step should investigate whether extreme events in ecological time series are driven by the intrinsic population dynamics or related to extreme events in physical forcing.

Acknowledgments

We thank the Service de Navigation de la Seine (SNS) for providing the Seine river flow data. We also thank the staff of the Laboratory of Oceanography of Villefranche for commitment to sampling and technical support, and particularly Serge Dallot whose efforts have made the *Centropages typicus* time-series possible. Financial support have been provided by the Seine Aval programme and the IFB "Institut Français de la Biodiversité". PhD fellowships are provided by the French alternance fellowship for Chinese doctoral students (YH) and a fellowship from the Gabon government (SZB).

References

Benzi, R., Ciliberto, S., Tripiccione, R., Baudet, C., Massaioli, F. and Succi, S. (1993) Extended self-similarity in turbulent flows. Phys. Rev. E 48, 29-32.

Caers, J., Beirlant, J. and Vynckier, P. (1998) Bootstrap confidence intervals for tail indices. Comput. Statist. Data Anal. 26, 259-277.

Coughlin, K. T. and Tung, K. K. (2004) Eleven year solar cycle in the lower stratosphere extracted by the empirical mode decomposition method. Adv. Space. Res 34, 323-329.

Darwin, C. (1859) *On the origin of the species by natural selection*. Murray, London.

Delechelle, E., Lemoine, J. and Niang, O. (2005) Empirical Mode Decomposition: An Analytical Approach for Sifting Process. IEEE Signal Process. Lett. 12, 764-767.

Elton, C. S. (1924) Periodic Fluctuations in the Numbers of Animals: Their Causes and Effects. J. Exp. Biol. 2, 119-163.

Flandrin, P. and Goncalves, P. (2005) Empirical Mode Decompositions as Data-Driven Wavelet-Like Expansions. Int. J. of Wavelets, Multires. and Info. Proc 2, 477–496.

Flandrin, P., Rilling, G. and Goncalves, P. (2004) Empirical mode decomposition as a filter bank. IEEE Signal Process. Lett. 11, 112-114.

Frisch, U. (1995) *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press, Cambridge.

Frontier, S., Pichod-Viale, D., Leprêtre, A., Davoult, D. and Luczak, C. (2004) *Ecosystèmes: structure, fonctionnement, évolution*. Dunod, Paris.

Hill, B. M. (1975) A Simple General Approach to Inference About the Tail of a Distribution. Ann. Statist. 3, 1163-1174.

Huang, N. E., Shen, Z. and Long, S. R. (1999) A new view of nonlinear water waves: The Hilbert Spectrum. Annu. Rev. Fluid Mech. 31, 417-457.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, E. H., Zheng, Q., Tung, C. C. and Liu, H. H. (1998) The empirical mode decomposition method and the Hilbert spectrum for non-stationary time series analysis. Proc. R. Soc. Lond. A 454, 903-995.

Hurst, H.E. (1951) Long-term storage capacity of reservoirs. Trans. Am. Soc. Civil Eng. 116, 770-799.

Hwang, P. A., Huang, N. E. and Wang, D. W. (2003) A note on analyzing nonlinear and nonstationary ocean wave data. Appl. Ocean Res. 25, 187-193.

Kantelhardt, J. W., Rybski, D., Zschiegner, S.A., Braun, P., Koscielny-Bunde, E., Livina, V., Havlin, S. and Bunde A. (2003) Multifractality of river runoff and precipitation: comparison of fluctuation analysis and wavelet method. Physica A 330, 240-245.

Katz, R. W., Brush, G. S. and Parlange, M. B. Statistics of extremes: modeling ecological disturbances. Ecology 86, 1124-1134.

Koscielny-Bunde, E., Kantelhardt, J.W., Braun, P., Bunde, A. and S. Havlin (2006). Long-Term Persistence and Multifractality of River Runoff Records: Detrended Fluctuation Studies. J. Hydrology 322, 120-137.

Leadbetter, M. R., Lindgren, G. and Rootzen, H. (1983) *Extremes and related properties of random sequences and processes*. Springer, Berlin.

Loutridis, S. J. (2005) Resonance identification in loudspeaker driver units: A comparison of techniques. Appl. Acoust. 66, 1399-1426.

Mitzenmacher, M. (2003) A brief history of generative models for power law and lognormal distributions. Internet Mathematics 1, 226-51.

Montanari, A., Rosso, R. and Taqqu, M.S. (1997) Fractionally differenced ARIMA models applied to hydrologic time series: identification, estimation and simulation. Water Resour. Res. 33, 1035-1044.

Newman, M. E. J. (2005) Power laws, Pareto distributions and Zipf's law. Contemp. Phys. 46, 323-351.

Ohman, M. D. and Hirche, H. J. (2001) Density-dependent mortality in an oceanic copepod population. Nature 412, 638-41.

Pandey, G., Lovejoy, S. and Schertzer, D. (1998) Multifractal analysis of daily diver flows including extremes for basins of five to two million square kilometres, one day to 75 days. J. Hydrology 208, 62-81.

Peng, L. (1998) Asymptotically unbiased estimators for the extreme-value index. Statist. Probab. Lett 38, 107-115.

Ponomarenko, V. I., Prokhorov, M. D., Bespyatov, A. B., Bodrov, M. B. and Gridnev, V. I. (2005) Deriving main rhythms of the human cardiovascular system from the heartbeat time series and detecting their synchronization. Chaos, Solitons and Fractals 23, 1429-1438.

Rilling, G., P. Flandrin and P. Goncalves (2003). On Empirical Mode Decomposition and Its Algorithms. *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*.

Salisbury, J. I. and Wimbush, M. (2002) Using modern time series analysis techniques to predict ENSO events from the SOI time series. Nonlin. Processes Geophys. 9, 341-345.

Schmitt, F., Lovejoy, S., and Schertzer, D. (1995) Multifractal analysis of the Greenland ice-core Project climate data. Geophys. Res. Lett. 22, 1689-1692.

Schmitt, F. and C. Nicolis (2002). Scaling of Return Times for a High-Resolution Rainfall Time Series. Fractals 10, 285-290.

Schmitt, F. G., Molinero, J.C. and Zongo, B. S. (2007) Nonlinear dynamics and intermittency in a long-term copepod time series. Comm. Nonlinear Sci. Num. Sim. (In Press).

Seuront, L. and Spilmont, N. (2002) Self-organized criticality in intertidal microphytobenthos patch patterns. Physica A 313, 513-539.

Tessier, Y., Lovejoy, S., Hubert, P., Schertzer, D. and Pecknold S. (1996) Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions. J. Geophys. Res. 101, 26427-26440.

Veltcheva, A. D. and Soares, C. G. (2004) Identification of the components of wave spectra by the Hilbert Huang transform method. Appl. Ocean Res. 26, 1-12.

Wu, Z. and Huang, N. (2004) A Study of the Characteristics of White Noise with the Empirical Mode Decomposition. Proc. R. Soc. Lond. A 460, 1597-1611.

16 The Arctic Ocean as a Coupled Oscillating System to the Forced 18.6 Year Lunar Gravity Cycle

Harald Yndestad

Aalesund University College, hy@hials.no

Abstract. The Arctic Ocean is a substantial energy sink for the Earth's Northern Hemisphere. Future fluctuations in its energy budget will have a major influence on the Arctic climate. A wavelet spectrum analysis of an extensive historical Arctic data series concludes that we may be able to understand Arctic climate dynamics as an oscillation system coupled to the forced 18.6 yr lunar nodal gravity cycle. This paper presents the results from a wavelet spectrum analysis of the data series which included polar movement, Arctic ice extent and the inflow of North Atlantic Water to the Norwegian Sea. The investigation shows a correlation better than R=0.6 between the astronomic 18.6 yr lunar nodal gravity cycle and identified 18 yr dominant cycles in the data series. The identified 18 yr cycles have phase-reversals synchronized to a 74 yr sub-harmonic lunar nodal cycle.

1 Introduction

The cold Arctic Ocean acts as an energy sink that influences the climate. The sun is never high in the sky, and most of its energy is reflected back to space by snow and ice in the summer. Cooled Arctic water accumulates in circulating layers. The Arctic Ocean has four layers, all with different densities and circulation patterns: Arctic Surface Water, Atlantic Water, Deep Water, and Bottom Water. Atlantic Water from the Norwegian Sea enters the Eurasian Basin through the Fram Strait underneath the surface layer. At a level of about 200–900 m, it follows the continental slope east and north until it meets the Transpolar Drift, it then returns to the Greenland Sea. Deep Water down to 2600 m from the Norwegian Sea is exchanged between the Greenland Sea and the Eurasian Basin. Deep Water residence time in the Eurasian Basin is estimated to be about 75 years (Bonisch and Schlosser, 1995). An oscillation disturbance in the Arctic system may introduce an oscillating effect on Arctic climate.

Most of the changes in Arctic climate are influenced by inflow of North Atlantic Water. North Atlantic Water passes through the Faroe-Shetland Channel and into the Norwegian Sea. The current continues northward with an inflow to the Barents Sea and to the Arctic Ocean through the Fram Strait. The current circulates in the Arctic Ocean and influences the Arctic ice extent before it returns to the Greenland Sea. From the Greenland Sea, the cold water continues south with the East Greenland Current and the Labrador Current. East of Newfoundland, the cold current meets the warm Atlantic drift from the south, and some water returns to the northern Atlantic Ocean.

The North Atlantic Current has large temperature fluctuations (Fig. 16.2), and it has been known for decades that this current has a major influence on the climate of northern Europe. A possible cause for the large temperature fluctuations in the North Atlantic Water is the tidal oscillations resulting from gravity effects between the earth, moon and sun (Yndestad, Turrell and Ozhigin 2004). A gravity effect between these three bodies results in a set of long orbital cycles that may introduce climate oscillations on the earth (Pettersson 1915; Currie 1984; Imbrie and Imbrie 1980; Satterley 1996; Yndestad 2006). An oscillating gravity effect on the climate fluctuations implies that there may be a coupled oscillation between gravity cycles and Arctic climate. This study investigates the relationship between the astronomic 18.6 yr lunar nodal tide and the dominant cycles in polar motion, North Atlantic Water and Arctic ice extent. The investigation concludes that the Arctic climate variability can be thought of as an oscillator coupled to the forced 18.6 yr lunar nodal gravity cycle. This coupled Arctic oscillation has temporary harmonic cycles in which interference between cycles may introduce phase-reversals.

2 Materials and Methods

The polar position data series (y-direction) is based on official data from the International Earth Rotation Service (IERS) and covers the years from 1846 until 2005. The time series can be found at: <http://hpiers.obspm.fr/eoppc/eop/eopc01/>. The time series contains 10 samples per year from 1846 to 1900, and 20 samples per year from 1900 to 2002, with the celestial pole offset represented in arc degrees.

North Atlantic Water is estimated within the core of the Slope Current on the Scottish side of the Faroe-Shetland Channel where the temperature and salinity at the standard depth show a maximum salinity (Turrell et al. 1999). The data series covers 1893 to 2002 and has no values in the periods from 1895 to 1902, 1915 to 1922, 1930 to 1933, and 1941 to 1946. In these periods the data are cubic spline interpolated. The data series is provided by the FRS Marine Laboratory, Aberdeen (Turrell, personal communication).

The Arctic ice extent data series represents the sum of the Greenland Sea and Barents Sea ice extents. The area of the Greenland Sea ice extent discussed in this paper covers the Greenland Sea, the Iceland Sea and the Norwegian Sea bounded by 30°W, 10°E and 80°N. The data are based on values taken in April from 1850 until 2000. The Barents Sea ice extent covers the Norwegian Sea, the Barents Sea and the Kara Sea bounded by 10°E, 80°N and 70°E. The data are based on values from April between 1850 and 2000 (Vinje 2001). The coupled oscillating system may be represented by the simple system model

$$S(t) = \{B(t), \{S_s(t), S_m(t), S_e(t)\}\} \quad , \tag{1}$$

where $S_s(t)$ represents the sun system, $S_m(t)$ the moon system, $S_e(t)$ the earth system and B(t) the mutual gravity relation between the planetary systems. The mutual relation has a set of long-term cycles that may influence dynamics on the earth system. The 18.6 yr lunar nodal amplitude cycle and the 18.6/2=9.3 yr lunar nodal phase cycle are both important gravity cycles. The astronomic lunar nodal amplitude cycle period of T=18.6134 years introduces a periodic tide that has both vertical and horizontal components. The vertical component has a global influence on sea level and its greatest influence is at the Equator and at the Arctic Ocean, while it has a minimal influence at about 30 degrees from the Equator. The 18.6 yr amplitude tide has a maximum in November 1987, which represents a phase delay of about $\varphi_T(t)=0.90\pi$ rad in Eq. 3. The vertical tide causes a horizontal current component that has a phase angle of about $\varphi_T(t)=(0.90\text{-}0.50)\pi$ rad. The 9.3 yr lunar nodal cycle phase tide has a period of $T/2$=9.3 years and a phase-angle of about $\varphi_{T/2}(t)=1.41\pi$ rad (Pugh 1996).

In this analysis, the lunar nodal gravity cycles represents a stable long-term forced oscillation on the coupled earth system [$S_e(t)$]. The coupled earth system is expected to respond by a chain of reactions related to the movement of the earth's axis, earth mantle dynamics and tidal oscillation in the sea systems. A lunar nodal tide influence on circulating water may be modelled by the state model:

$$x(t) = a \cdot x(t-\tau) + u(t)\,, \tag{2}$$

where u(t) represents the tidal input to the circulation, x(t) the state at the time t, τ the ocean circulation delay and a the circulation temperature loss. Energy from a forced oscillating lunar nodal tide into a positive circulating water feedback system will cause energy to be accumulated in a set of sub-harmonic cycles:

$$x(t) = \sum_k a_k(t)\sin(k\omega_T t + \varphi_{kT}(t)) + v(t)\,, \tag{3}$$

where $v(t)$ is a disturbance from an unknown source, $a_k(t)$ represents the cycle amplitude, $\omega_T=2\pi/T$ (rad y^{-1}) the cycle period, and $\varphi_{kT}(t)$ is the time-dependent phase angle. A cycle number k may have values k=1,2,3… on harmonic cycles and k=1/2,1/3… on sub-harmonic cycles in the lunar nodal spectrum, as evaluated for times t=1900,1901,…,2005. The long-term analysis was based on the assumption that there is a possible 74 yr sub-harmonic cycle in the Arctic data series where the identified phase angle is $\varphi_{4T}(t)=0.29\pi$ rad (Yndestad 2006). A sub-harmonic cycle $a_{kT}(t)\sin(\omega_{kT}t)$ may have accumulated an amplitude $a_{kT}(t) >> a_T(t)$. This may introduce a phase-reversal in the forced lunar nodal cycle $a_{kT}(t)\sin(\omega_{kT}t)$ by interference. This investigation identified a phase-reversal in the 18.6 yr lunar nodal cycle which is related to the 74 yr sub-harmonic cycle.

Phase-reversals in the cycle periods introduce a time-variant stochastic and chaotic process. This property excludes traditional spectrum analysis methods to identify cycle periods and cycle phase. In this study, the time-series have been analyzed by wavelet transformation to identify the dominant cycle periods $u_k(t)$ and the time-variant

phase angle $\varphi_{kT}(t)$. The periodicity was identified by a three-step investigation. The first step was to compute the wavelet spectrum by the transformation:

$$W_{a,b}(t) = \frac{1}{\sqrt{a}} \int_R x(t) \Psi(\tfrac{t-b}{a}) \mathrm{d}t \,, \qquad (4)$$

where $x(t)$ is the time-series analyzed, and $\Psi(\ )$ is a coiflet wavelet impulse function (Daubechies 1992; Matlab 1997). $W_{a,b}(t)$ is a set of wavelet cycles, b is the translation in time, and a is the time-scaling parameter in wavelet transformation. The relationship between the wavelet scaling (a) and the sinus period T is about $T \approx 1.2a$. In this analysis, the translation is b=0, so the computed wavelet transformation $W_a(t)$ represents a moving correlation between $x(t)$ and the impulse function $\Psi(\ )$ over the whole time-series $x(t)$.

Using this wavelet transformation, it is possible to identify single, long-period cycles in a short time-series. Errors at the beginning and at the end of a time-series are reduced in the following manner. The time-series are scaled in amplitude and to the zero mean value by the scaling transformation $y(t)=[x(t)–\text{mean}(x(t))]/\text{var}(x(t))$, where $x(t)$ is the time-series, and $y(t)$ is the scaled time-series. Subsequently, the time-series is expanded with symmetric values at the beginning and end of the time-series. The cycle periods of single dominant-wavelet cycles are identified by computing the autocorrelation for the wavelet spectrum $R_w(\tau)=\mathrm{E}[W_a(t)W_a(t+\tau)]$. Dominant, stationary wavelet cycles have maximum values in the autocorrelation functions. Periodic cycles in the autocorrelation function of the wavelet spectrum demonstrate that there is a stationary cycle. The cycle period phase is identified by the optimum correlation between dominant wavelets and lunar nodal cycles by $R_{kT}(\tau)=E[W_{ak}(t)u_{kT}(t)]$, where $u_{kT}(t)$ is a kT lunar nodal cycle period and the phase angle $\varphi_{kT}(t)$ is a free variable. The correlation quality is computed by the Pearson correlation coefficient $Q_{kT}= R*\text{sqr}[(n-2)/(1- R*R)] \sim t(n-2)$ where R is the correlation coefficient and n is the number of samples. Q_{kT} is tested against a t-distribution by (n–2) numbers of freedom to determine whether the correlation is statistically significant (Storch and Zwiers, 1999).

3 Results

3.1 Polar position

The pole position was influenced by the constellation between the earth, sun and the moon. The position changed in the y-direction from about 0.1 to 0.3 arc degrees, representing a displacement of approximately 5 to 15 meters. A computed wavelet spectrum from the polar position time series had dominant cycles with periods of about 1.2, 6, 18 and 74 years (Fig. 16.1). The correlations with the lunar nodal harmonic cycles of 18.6/3=6.3, 18.6 and 18.6*4=74.4 were estimated, with $R_{T/3}$=0.44, R_T=0.7 and R_{4T}=0.86. The Pearson correlation coefficients are estimated to $Q_{T/3}$ =6.2, Q_T =5.2, Q_{4T} = 20.5 which gives a better than 95% confidence in a t-distribution, or

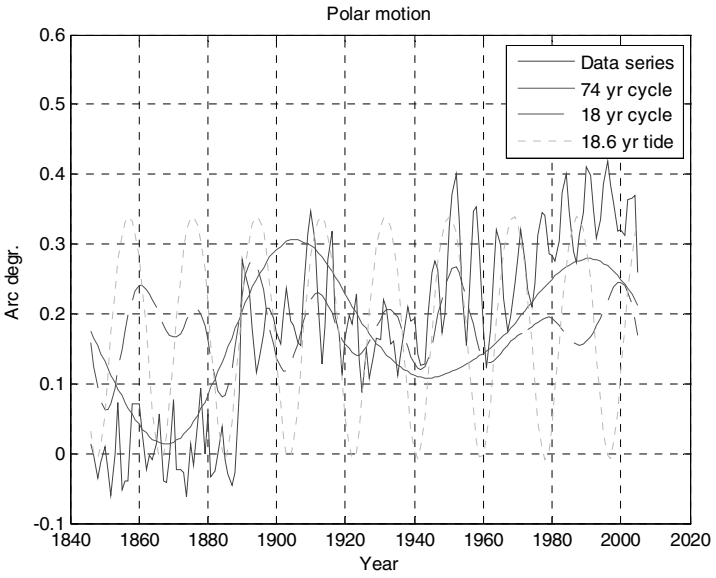


Fig. 16.1. Polar motion in the y-direction relative to the dominant 18 yr wavelet cycle, the 18.6 yr lunar nodal tide cycle and the dominant 74 yr wavelet cycle. (A color version of this figure appears between pages 196 and 197).

t(78)>0.95. (Yndestad 2006). The 1.2 yr (Chandler wave) was most likely a 5 harmonic cycle from the forced 6.2 yr harmonic nodal cycle.

The dominant 18 yr cycle had an estimated phase angle [$\varphi_{18}(t)$] of 0.90π rad between 1850 and 1960, which is the same as the lunar nodal cycle. After approximately 1960, the 18 yr cycle showed a phase-reversal. The dominant 74 yr wavelet cycle was at a minimum at about 1970 and 1940 and at a maximum at about 1910 and 1990, and turning points were observed at about 1890 and 1925. The phase-reversal of the 18 yr cycle introduced a phase delay of about 18.6/2 years and the next turning point was in 1970. The cycle phase angle was estimated to be $\varphi_{74}(t)$=1.29π rad.

3.2 North Atlantic water

Figure 16.2 shows the time-series of the North Atlantic Water temperature on the Scotland side of the Faroe-Shetland Channel from 1900 to 2005 and represents one of the longest oceanographic time-series in the world. The data series demonstrates that the temperature has large fluctuations within a period of one hundred years. A wavelet

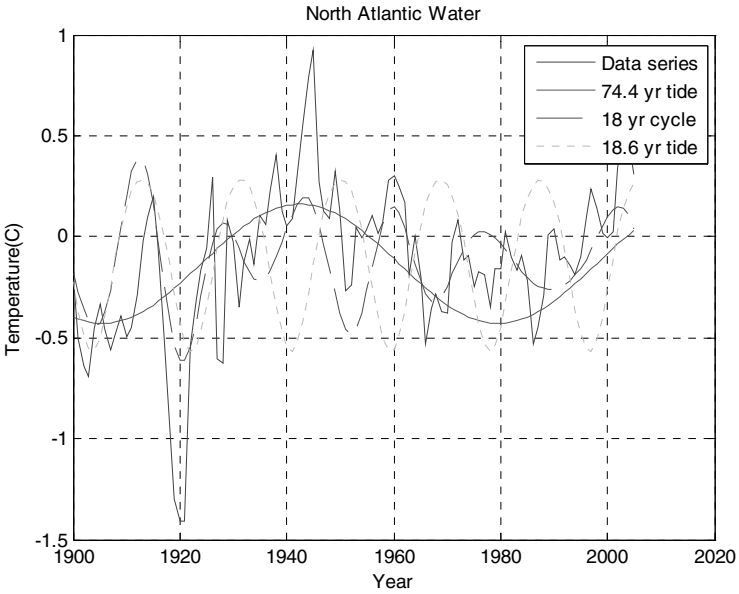


Fig. 16.2. NAW temperature data series shown alongside the 74 yr harmonic lunar nodal tide, the identified 18 yr wavelet cycle, and the astronomic 18.6 yr lunar nodal tide cycle. (A color version of this figure appears between pages 196 and 197).

analysis of the data series identified dominant cycles that correlated with lunar nodal cycles of about 9, 18 and 74 years (Fig. 16.2). The dominant lunar nodal cycle periods correlated with the astronomic lunar nodal tides, but had an unstable phase. The identified 74 yr wavelet cycle, $W_{74}(t)$, had an estimated phase of about $\varphi_{74}(t)=(0.29-0.10)\pi$ rad and represents the long-term mean temperature fluctuation. The correlation is estimated to $R_{4T}=0.93$, $Q_{4T}= 27.8$ and N=104 to the sub-harmonic cycle $u_{4T}(t)$. The 74 yr cycle had a reversed phase compared to the 74 yr polar movement cycle. The estimated phase of the identified 18 yr wavelet cycle $W_{18}(t)$ was about $\varphi_{18}(t)=0.90\pi$ rad for the period from 1900 to 1925, the same phase angle as the lunar nodal tide. At about 1925, the 74 yr period in polar movement changed to a negative state, the Atlantic water temperature changed to a positive state and the 18 yr cycle had a phase-reversal from 1925 to 2004. In this period the correlation to the astronomic 18.6 yr lunar tide was estimated to be $R_T=-0.68$, $Q_T=8.3$ and N=80 samples.

3.3 Arctic ice extent

A wavelet analysis of the Arctic ice extent data series identified the same dominant cycles of about 18 and 74 years as identified in the Polar motion data series and the North Atlantic Water. The dominant lunar nodal cycle periods correlated with the

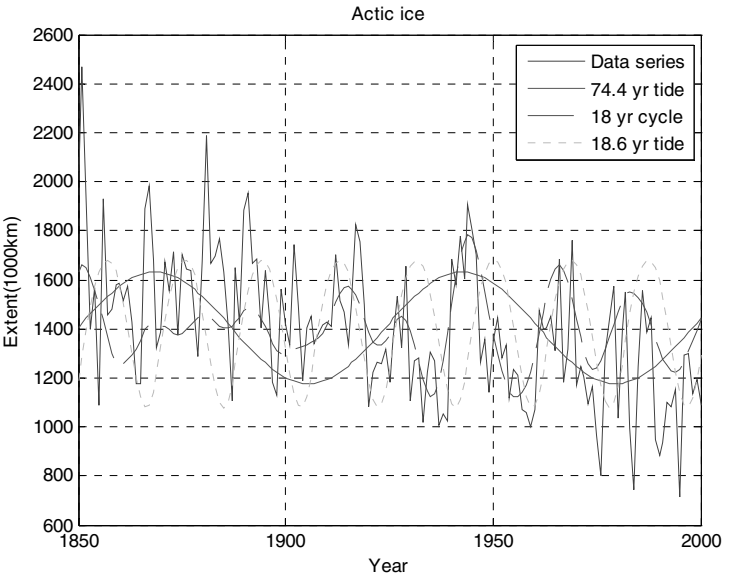


Fig. 16.3. The Arctic ice extent data series in relation to the 74 yr harmonic lunar nodal tide, the identified 18 yr wavelet cycle and the astronomic 18.6 yr lunar nodal tide cycle. (A color version of this figure appears between pages 196 and 197).

astronomic lunar nodal tides, but had an unstable phase. The identified 74 yr wavelet cycle, $W_{74}(t)$, had a correlation of R_{4T}=0.91, Q_{4T}=26 in N=149 samples with the 74.4 yr lunar nodal cycle and an estimated phase, $\varphi_{74}(t)$, of (0.29-0.18)π rad, representing a delay from North Atlantic Water of about 7 years.

The identified 18 yr wavelet cycle, $W_{18}(t)$, had an estimated phase of about $\varphi_{18}(t)$=(0.90+0.5)π rad and R_T=0.4 during the period from 1850 to about 1925 when the 74 yr cycle was becoming positive. From about 1925 to 2000, the 18 yr cycle had a phase-reversal and changed to the phase $\varphi_{18}(t)$=(0.90-0.5)π rad. In this period, the correlation to the astronomic 18.6 yr lunar tide was estimated to be R_T=0.84 Q_T=17 in N=75 samples. The phase delay of 0.5π rad corresponded to the lunar nodal cycle current and could be explained by a delay in the circulation of Arctic water.

4 Discussion

The wavelet spectrum analysis shows that polar motion, North Atlantic Water temperature and Arctic ice extent have temporary dominant fluctuations and have a correlation better than R=0.68 with harmonics from the astronomic 18.6 yr lunar nodal cycle. The exception is the first part of the Arctic ice data when there were unreliable observations (Vinje 2001). Polar motion time series had temporary dominant cycles of about 18.6/15, 18.6/6, 18.6 and 4*18.6 years, the North Atlantic Water at about 18.6/2, 18.6 and 4*18.6 years and Arctic ice extent at about 18.6 and 4*18.6

years. The dominant 74 yr long-term fluctuations correlated with a sub-harmonic cycle of 18.6*4=74.4 years. The North Atlantic Water 74 yr cycle had the same phase as the 74 yr cycle of Arctic ice extent and reversed phase related to the 74 yr cycle in polar motion. The mean Arctic ice extent has been decreasing since about 1825, when ice extent was maximal. This indicates that longer sub-harmonic cycles influencing climate fluctuations may exist.

The 18 yr cycles had phase-reversals at about 1925 and 1960, when the 74 yr sub-harmonic cycles shifted in a new direction. This phase-reversal introduces a time variant stochastic process in the time series and explains why the lunar nodal spectrum has been difficult to identify by traditional spectrum analysis methods. The wavelet transform represents a linear filter that separates periodic fluctuations in the time series. The dominant cycles are identified by a dominant periodic pattern in the wavelet autocorrelation spectrum. Dominant periodic fluctuations may be studied in separated time windows. The analysis method has been developed from many trials of test data. It is, however, the coherence between the identified cycle periods and phase in the many time series from nature that has demonstrated the robust nature of this wavelet analysis.

The 18.6 yr earth axis nutation movement was discovered by James Bradley (1693-1762). The influence of the 18.6 yr lunar nodal tide on climate has been discussed since the work of G. H. Darwin (1880) and Pettersson (1915). Russian scientists investigated this relationship up to the 1960's (Maksimov and Smirnov 1965, 1967). Currie (1987) identified the lunar nodal cycle in a number data series, and Keeling and Worf (1997), (Yndestad, Turrell and Ozhigin 2004; Yndestad 2006) predicted a set of strong 18-yr tidal waves in the Atlantic Ocean.

The lunar nodal tide represents only about 5% of the daily diurnal tide from the moon. A small lunar nodal tide has, however, a lot of power when integrated in time and space over a period of 9 years. A possible mechanism directly linking the long-term tides to the oceanic thermohaline circulation in the North Atlantic and Nordic Seas is the role that oceanic tidal mixing may play in resupplying the potential energy required by the circulation (Munk and Wunsch, 1998; Wunsch and Ferrari, 2004). The budget of mechanical energy input to the ocean is not well understood. Tidal mixing and polar movement may well be an important controlling mechanism on the properties, and perhaps magnitude, of the thermohaline circulation. Deep Water residence time in the Eurasian Basin of the Arctic Ocean is estimated to be about 75 years (Bonisch and Schlosser, 1995). The amplitude and phase of this circulation may have been influenced by the 18.6 tide and the polar motion. The 74 yr cycle of warm North Atlantic Water and cold Arctic ice extent have about the same phase. In this analysis, Arctic ice extent represents the sum of Barents Sea ice extent and Greenland Sea ice extent. The Barents Sea ice extent 74 yr cycle has a phase angle of about 0.25π rad. In the Greenland Sea, the 74 yr cycle is delayed to about -0.22π rad (Yndestad 2006). This indicates a relation between long-term Arctic ice fluctuations and long-term changes North Atlantic Water. The cause of the estimated phase-reversals is unclear. A possible explanation is interference between the 18.6 yr tide and the 74 yr sub-harmonic cycle.

5 Conclusions

The wavelet analysis identified dominant Arctic climate oscillations that correlated with the 18.6 yr lunar nodal gravity cycle in Polar movement, North Atlantic Water and Arctic ice extent over the past 100 years. The synchronization of cycle periods and phase indicates that the Arctic Ocean system oscillations are coupled to the forced 18.6 yr gravity cycle of the moon. The phase-reversals in the 18 yr tide are explained by interference between a 74 yr sub-harmonic cycle and the shorter 18 yr cycle. This phase-reversal property introduces a temporary time-variant system that may be understood as chaotic behaviour.

References

Bonisch, G and Schlosser, P. 1995. Deep water formation and exchange rates in the Greenland/ Norwegian Seas and the Eurasian Bain of the Arctic Ocean derived from tracer balances. Prog. Oceanography. 35:29-52.

Currie, R. G.1984. Evidence for 18.6 year (*sic*) lunar nodal (*sic*) drought in western North America during the past millennium: Journal of Geophysical Research.89:1295-1308.

Currie, R. G. 1987. Examples and implications of 18.6- and 11--year terms in world weather records, Chap. 22, p. 378-403 *in* M.R. Rampino, J.E. Sanders, W.S. Newman,and L.K. Konigsson, *eds*. Climate: History, periodicity, and predictability: International Symposium held at Barnard College, Columbia University, New York, New York, 21-23 May 1984 (R. W. Fairbridge Festschrift): New York, NY, Van Nostrand Reinhold Publishing Corp., 588 p.

Darwin, G.H. 1880. On the secular change of the orbit of a satellite revolving about a tidally distorted planet. Philosophical Transactions of the Royal Society of London. 171: 713-891.

Daubechies I. 1992. Ten lectures of wavelet. SIAM Journal on Mathematical Analysis. 24:499-519.

Imbrie John, Imbrie John Z. 1980. Modelling the Climate response to Orbital Variations. Science. 207, 29 February 1980.

Keeling, Charles D. and T. P. Whorf. 1997. Possible forcing global temperature by oceanic tides. Proceedings, National Academy of Sciences of the United States. 94:8321-8328.

Maksimov, I. V. and N.P. Smirnov. 1965. A contribution to the study of causes of long-period variations in the activity of the Gulf Stream. Oceanology. 5:15-24.

Maksimov, I. V. and N.P. Smirnov. 1967. A long-term circumpolar tide and its significance for the circulation of ocean and atmosphere. Oceanology 7: 173-178.

Matlab. 1997. Matlab. Wavelet Toolbox. Users Guide. The Math Works Inc.

Munk, W; Wunsch, C (1998). Abyssal recipes II: energetics of tidal and wind mixing

Pettersson, Otto, 1915, Long periodical (*sic*) variations of the tide-generating force: Conseil Permanente International pour l'Exploration de la Mer (Copenhagen), Pub. Circ. No. 65, p. 2-23.

Pugh, D T. 1996. Tides, Surges and Mean Sea-Level. John Wiley & Sons. New York.

Satterley, A. K. 1996. The Milankovitch Theory. Earth-Science Reviews. 40:181-207.

Storch, Hans von and Zwiers Francis W. 1999: Statistical Analysis in Climate Research. Cambridge University Press. ISBN 0 521 01230 9.

Turrell, WR, Sherwin, TJ, Jeans, DRG, et al. 1999. Eddies and mesoscale deflection of the slope current in the Faroe-Shetland Channel. DEEP-SEA RESEARCH PT I 46(3):415-438 MAR 1999.

Vinje, Torgny. 2001. Anomalies and trends of sea ice extent and atmospheric circulation in the Nordic Seas during the period 1864-1998. Journal of Climate. 14:255-267.

Wunsch C and Ferrari, R (2004). Vertical mixing, energy, and the general circulation of the oceans. Annual Review of Fluid Mechanics, Vol. 36: 281-314.

Yndestad H, Turrell W R, Ozhigin V. 2004. Temporal linkages between the Faroe-Shetland time series and the Kola section time series. ICES Annual Science Conference. Sept 2004. Vigo. ICES CM 2004/M.

Yndestad, H: 2006. The influence of the lunar nodal cycle on Arctic climate. Journal of Marine Science. 63. 401-420.

17 Dynamical synchronization of truth and model as an approach to data assimilation, parameter estimation, and model learning

Gregory S. Duane and Joseph J. Tribbia

National Center for Atmospheric Research, Boulder, Colorado,
gduane@ucar.edu

1 Synchronized Chaos and Data Assimilation

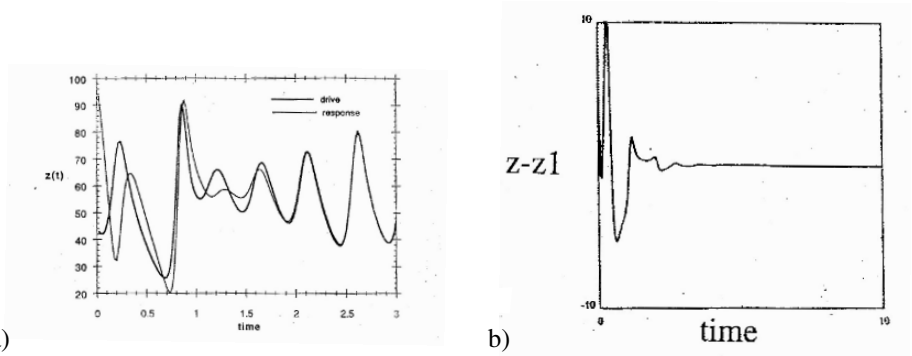
1.1 Synchronized chaos

Synchronization of weakly coupled oscillators has been known since the time of Huygens (1673), who observed that pendulum clocks hung on a common wall tend to fall into antisynchronized motion. Synchronization of regular oscillators with limit cycle attractors is in fact ubiquitous in Nature (Strogatz 2003). Only recently however, has the synchronization of chaotic oscillators become known. The phenomenon was first brought to light by Fujisaka and Yamada (1983) and independently by Afraimovich, Verichev, and Rabinovich (1987), but extensive research on the subject in the '90s was spurred by the seminal work of Pecora and Carroll (1991), who considered configurations such as the following combination of Lorenz systems:

$$\begin{aligned} \dot{X} &= \sigma(Y - X) \\ \dot{Y} &= \rho X - Y - XZ \qquad\qquad \dot{Y}_1 = \rho X - Y_1 - XZ_1 \\ \dot{Z} &= -\beta Z + XY \qquad\qquad \dot{Z}_1 = -\beta Z_1 + XY_1 \end{aligned} \tag{1}$$

which synchronizes rapidly, slaving the Y_1, Z_1-subsystem to the master X, Y, Z-subsystem, as seen in Fig. 17.1, despites differing initial conditions and despite sensitive dependence on initial conditions.

If we imagine that the first Lorenz system represents the world, and that the second Lorenz system is a predictive model, then synchronization effects *data assimilation* (Daley 1991) of observed variables into the running model. The only observed variable in the foregoing example is X, but that is sufficient to cause the desired convergence of model to truth. Synchronization is known to be tolerant of reasonable levels of noise, as might arise in the observation channel, and occurs with partial coupling schemes that do not completely replace a model variable with a variable of the observed system.



a) b)

Fig. 17.1. The trajectories of the synchronously coupled Lorenz systems in the Pecora-Carroll complete replacement scheme (1) rapidly converge (a). Differences between corresponding variables approach zero (b).

Specifically, systems can also synchronize when coupled *diffusively*, as with a pair of bidirectionally coupled Rossler systems:

$$\begin{aligned} \dot{X} &= -Y - Z + \alpha(X_1 - X) & \dot{X}_1 &= -Y_1 - Z_1 + \alpha(X - X_1) \\ \dot{Y} &= X + aY & \dot{Y}_1 &= X_1 + aY_1 \\ \dot{Z} &= b + Z(X - c) & \dot{Z}_1 &= b + Z_1(X_1 - c) \end{aligned} \tag{2}$$

where α parametrizes the coupling strength. Contrary to naive expectations, the tendency to synchronize does not increase monotonically with α, but synchronization will indeed occur for sufficiently large α. The diffusive coupling also works unidirectionally.

For a pair of coupled systems that are not identical, synchronization may still occur, but the correspondence between the states of the two systems in the synchronized regime is different from the identity. In this situation, known as *generalized synchronization*, we have two different dynamical systems

$$\dot{\mathbf{x}} = F(\mathbf{x}) \tag{3}$$

$$\dot{\mathbf{y}} = G(\mathbf{y}) \tag{4}$$

with $\mathbf{x} \in R^N$ and $\mathbf{y} \in R^N$. If the dynamics are modified so as to couple the systems:

$$\dot{\mathbf{x}} = \hat{F}(\mathbf{x}, \mathbf{y}) \tag{5}$$

$$\dot{\mathbf{y}} = \hat{G}(\mathbf{y}, \mathbf{x}) \tag{6}$$

the systems are said to be generally synchronized iff there is some invertible function $\Phi : R^N \rightarrow R^N$ such that $||\Phi(\mathbf{x}) - \mathbf{y}|| \rightarrow 0$ as $t \rightarrow \infty$. Identical synchronization may be transformed to generalized synchronization simply by a change of variables

in one system, but not the other, i.e. a change in the description of one of the systems (Rulkov, Sushchik, and Tsimring 1995). In this situation, the correspondence function Φ is known *a priori*. Generalized synchronization may be difficult to detect without prior knowledge of Φ.

Synchronization reduces the effective dimension of the phase space by half. Once synchronized, the 6-dimensional system (2) evolves in an 3-dimensional hyperplane, and is further constrained, as $t \rightarrow \infty$, to evolve on a strange attractor within that hyperplane. (For the coupled Lorenz systems (1), an auxiliary variable X_1 could be added, satisfying $\dot{X}_1 = \sigma(Y_1 - X_1)$, to define a system that is 6-dimensional before synchronization.) With *generalized synchronization* of nonidentical systems, the hyperplane becomes a *synchronization manifold* defined by an invertible correspondence function $\Phi : R^N \rightarrow R^N$. The N-dimensional manifold in $2N$-dimensional space is $\mathcal{M} \equiv \{(p, \Phi(p)) | p \in R^N\}$. The synchronization manifold is dynamically invariant: If $x(t)$ is a trajectory of a system such as (1) or (2), for $x \in R^{2N}$, and $x(t_1) \in \mathcal{M}$, then $x(t_2) \in \mathcal{M}$ for all $t_2 > t_1$. That is, a perfectly synchronized system remains synchronized.

It is commonly not the existence, but the stability of the synchronization manifold that distinguishes coupled systems exhibiting synchronization from those that do not (such as (2) for different values of α). N Lyapunov exponents can be defined for perturbations in the N-dimensional space that is transverse to the synchronization manifold $\mathcal{M}$. If the largest of these, $h^{\perp}_{max}$, is negative, then motion in the synchronization manifold is stable against transverse perturbations. In that case, the coupled systems will synchronize for some range of differing initial conditions. However, since $h^{\perp}_{max}$ only determines *local* stability properties, the size of the basin of attraction for the synchronized regime remains unknown. As $h^{\perp}_{max}$ is increased through zero, the system undergoes a *blowout bifurcation*. For small positive values of $h^{\perp}_{max}$, on-off synchronization occurs (a special case of on-off intermittency), as illustrated in Fig. 17.2b, where degradation results from a time lag in the coupling.

Synchronization is surprisingly easy to arrange, occurring for a wide range of coupling types. Synchronization degrades through on-off intermittency or through generalized synchronization. In the former case, vestiges of synchronization are discernible even far from the blowout bifurcation point (Duane 1997) as in Fig. 17.2c. Generalized synchronization is known to occur even when the systems are very different, as in the case of a Lorenz system diffusively coupled to a Rossler system. The two systems with attractors of different dimension are known to synchronize, but the correspondence function is not smooth (Pecora, Carroll, Johnson, Mar, and Heagy 1997).

The phenomenon of chaos synchronization is not restricted to low-dimensional systems. It is known, for instance, that two D-dimensional Generalized Rossler systems (each equivalent to a Rossler system for $D = 3$) will synchronize for any D, no matter how large, when coupled via only one of the D variables:

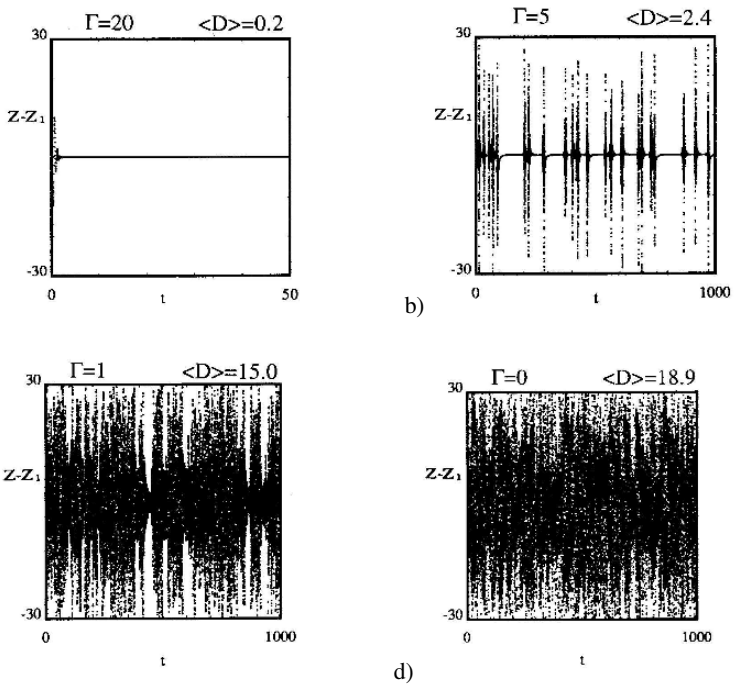


Fig. 17.2. The difference between the simultaneous states of two Lorenz systems with time-lagged coupling, represented by $Z(t) - Z_1(t)$ vs. t for various values of the inverse time-lag Γ illustrating complete synchronization (a), intermittent or "on-off" synchronization (b), partial synchronization (c), and de-coupled systems (d). Average euclidean distance $\langle D \rangle$ between the states of the two systems in X, Y, Z-space is also shown. The trajectories are generated by adaptive Runge-Kutta numerical integrations with $\sigma = 10.$, $\rho = 28.$, and $\beta = 8/3$.

$$\begin{aligned} \dot{x}_1^A &= -x_2^A + \alpha x_1^A + x_1^B - x_1^A & \dot{x}_1^B &= -x_2^B + \alpha x_1^B + x_1^A - x_1^B \\ \dot{x}_i^A &= x_{i-1}^A - x_{i+1}^A & \dot{x}_i^B &= x_{i-1}^B - x_{i+1}^B \qquad i = 2 \ldots D-1 \\ \dot{x}_D^A &= \epsilon + \beta x_D^A (x_{D-1}^A - d) & \dot{x}_D^B &= \epsilon + \beta x_D^B (x_{D-1}^B - d) \end{aligned} \tag{7}$$

Each system has an attractor of dimension $\approx D - 1$, for D greater than about 40, and a large number of positive Lyapunov exponents that increases with D.

The existence of synchronized chaos in naturally occuring systems was made more plausible by demonstrations of synchronization in spatially extended systems governed by PDE's (Kocarev, Tasev, and Parlitz 1997) Synchronization in geophysical fluid models was demonstrated by Duane and Tribbia (2001), originally with a view toward predicting and explaining new families of long-range teleconnections (Duane and Tribbia, 2004). Their findings are discussed in Section 3.2.

1.2 Synchronization-based data assimilation and Jungian synchronicity

Since the problem of data assimilation arises in any situation requiring a computational model of a parallel physical process to track that process as accurately as possible based on limited input, it is suggested here that the broadest view of data assimilation is that of machine perception by an artificially intelligent system. Like a data assimilation system, the human mind forms a model of reality that functions well, despite limited sensory input, and one would like to impart such an ability to the computational model. In the artificial intelligence view of data assimilation, the additional issue of model error can be approached naturally as a problem of machine learning, as discussed in the concluding section.

In this more general context, the role of synchronism is reminiscent of the psychologist Carl Jung's notion of synchronicity in his view of the relationship between mind and the material world. Jung had noted uncanny coincidences or "synchronicities" between mental and physical phenomena. In collaboration with Wolfgang Pauli (Jung and Pauli, 1955), he took such relationships to reflect a new kind of order connecting the two realms. The new order was taken to explain relationships between seemingly unconnected phenomena in the objective world as well, exising alongside the familiar order based on causality (Fig. 17.3). It was important to Jung and Pauli that synchronicities themselves were distinct, isolated events, but as described in Sect. 1.1, such phenomena can emerge naturally as a degraded form of chaos synchronization.

A principal question that is addressed in this chapter is whether the synchronization view of data assimilation is merely an appealing reformulation of standard treatments, or is different in substance. The first point to be made is that all standard data assimilation approaches, if successful, do achieve synchronization, so that synchronization defines a more general family of algorithms that includes the standard ones. It remains to determine whether there are synchronization schemes that lead to faster convergence than the standard data assimilation algorithms. It has been shown analytically that optimal synchronization is equivalent to Kalman filtering when the dynamics change slowly in phase space, so that the same linear approximation is valid at each point in time for the real dynamical system and its model. When the dynamics change rapidly, as in the vicinity of a regime transition, one must consider the full nonlinear equations and there are better synchronization strategies than the one given by Kalman filtering or ensemble Kalman filtering. The deficiencies of the standard methods, which are well

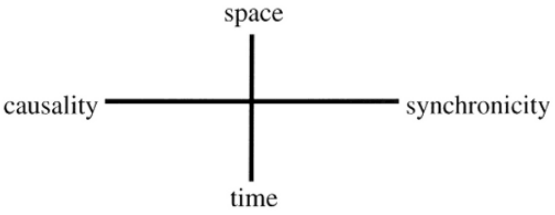


Fig. 17.3. Diagram constructed by Carl Jung, later modified by Wolfgang Pauli, to suggest relationships based on synchronicity as an "acausal connecting principle", existing alongside causal relationships. (Jung and Pauli, 1955; Peat, 1987)

known in such situations, are usually remedied by ad hoc corrections, such as "covariance inflation" (Anderson, 2001). In the synchronization view, such corrections can be derived from first principles.

2 Synchronization vs. conventional data assimilation in the nonlinear realm

2.1 Optimal coupling for synchronization

To compare synchronization to standard data assimilation, Duane, Tribbia, and Weiss (2006) inquired as to the coupling that is optimal for synchronization, so that this coupling could be compared to the gain matrix used in the standard 3dVar and Kalman filtering schemes. Their findings are reproduced here.

The general form of coupling of truth to model that we consider in this section is given by a system of stochastic differential equations:

$$\begin{aligned} \dot{\mathbf{x}}_T &= f(\mathbf{x}_T) \\ \dot{\mathbf{x}}_B &= f(\mathbf{x}_B) + \mathbf{C}(\mathbf{x}_T - \mathbf{x}_B + \boldsymbol{\xi}) \end{aligned} \tag{8}$$

where true state $\mathbf{x}_T \in R^n$ and the model state $\mathbf{x}_B \in R^n$ evolve according to the same dynamics, given by f, and where the noise $\boldsymbol{\xi}$ in the coupling (observation) channel is the only source of stochasticity. The form (8) is meant to include dynamics f described by partial differential equations, as in the last section. The system is assumed to reach an equilibrium probability distribution, centered on the synchronization manifold $\mathbf{x}_B = \mathbf{x}_T$. The goal is to choose a time-dependent matrix $\mathbf{C}$ so as to minimize the spread of the distribution.

Note that if $\mathbf{C}$ is a projection matrix, or a multiple of the identity, then Eq. (8) effects a form of nudging. But for arbitrary $\mathbf{C}$, the scheme is much more general. Indeed, continuous-time generalizations of 3DVar and Kalman filtering can be put in the form (8).

Let us assume that the dynamics vary slowly in state space, so that the Jacobian $\mathbf{F} \equiv Df$, at a given instant, is the same for the two systems

$$Df(\mathbf{x}_B) = Df(\mathbf{x}_T) \tag{9}$$

where terms of $O(\mathbf{x}_B - \mathbf{x}_T)$ are ignored. Then the difference between the two Eqs. (8), in a linearized approximation, is

$$\dot{\mathbf{e}} = \mathbf{F}\mathbf{e} - \mathbf{C}\mathbf{e} + \mathbf{C}\boldsymbol{\xi} \tag{10}$$

where $\mathbf{e} \equiv \mathbf{x}_B - \mathbf{x}_T$ is the synchronization error.

The stochastic differential equation (10) implies a deterministic partial differential equation, the Fokker-Planck equation, for the probability distribution $\rho(\mathbf{e})$:

$$\frac{\partial \rho}{\partial t} + \nabla_{\mathbf{e}} \cdot [\rho(\mathbf{F} - \mathbf{C})\mathbf{e}] = \frac{1}{2}\delta \nabla_{\mathbf{e}} \cdot (\mathbf{C}\mathbf{R}\mathbf{C}^T \nabla_{\mathbf{e}} \rho) \tag{11}$$

where $\mathbf{R}=<\boldsymbol{\xi}\boldsymbol{\xi}^T>$ is the observation error covariance matrix, and δ is a time-scale characteristic of the noise, analogous to the discrete time between molecular kicks in a Brownian motion process that is represented as a continuous process in Einstein's well known treatment. Equation (11) states that the local change in ρ is given by the divergence of a probability current $\rho(\mathbf{F}-\mathbf{C})\mathbf{e}$ except for random "kicks" due to the stochastic term.

The PDF can be taken to have the Gaussian form $\rho=N\exp(-\mathbf{e}^T\mathbf{K}\mathbf{e})$, where the matrix $\mathbf{K}$ is the inverse spread, and N is a normalization factor, chosen so that $\int \rho d^n\mathbf{e}=1$. For background error covariance $\mathbf{B}$, $\mathbf{K}=(2\mathbf{B})^{-1}$. In the one-dimensional case, $n=1$, where C and K are scalars, substitution of the Gaussian form in Eq. (11), for the stationary case where $\partial\rho/\partial t=0$ yields:

$$2B(C-F)=\delta RC^2 \tag{12}$$

Solving $dB/dC=0$, it is readily seen that B is minimized (K is maximized) when $C=2F=(1/\delta)B/R$.

In the multidimensional case, $n>1$, the relation (12) generalizes to the *fluctuation-dissipation relation*

$$\mathbf{B}(\mathbf{C}-\mathbf{F})^T+(\mathbf{C}-\mathbf{F})\mathbf{B}=\delta\mathbf{C}\mathbf{R}\mathbf{C}^T \tag{13}$$

that can be obtained directly from the stochastic differential equation (10) by a standard proof. $\mathbf{B}$ can then be minimized element-wise. Differentiating the matrix equation (13) with respect to the elements of $\mathbf{C}$, we find

$$\begin{aligned} &d\mathbf{B}(\mathbf{C}-\mathbf{F})^T+\mathbf{B}(d\mathbf{C})^T+(d\mathbf{C})\mathbf{B}+(\mathbf{C}-\mathbf{F})d\mathbf{B}\\ &=\delta[(d\mathbf{C})\mathbf{R}\mathbf{C}^T+\mathbf{C}\mathbf{R}(d\mathbf{C})^T] \end{aligned} \tag{14}$$

where the matrix $d\mathbf{C}$ represents a set of arbitrary increments in the elements of $\mathbf{C}$, and the matrix $d\mathbf{B}$ represents the resulting increments in the elements of $\mathbf{B}$. Setting $d\mathbf{B}=0$, we have

$$[\mathbf{B}-\delta\mathbf{C}\mathbf{R}](d\mathbf{C})^T+(d\mathbf{C})[\mathbf{B}-\delta\mathbf{R}\mathbf{C}^T]=0 \tag{15}$$

Since the matrices $\mathbf{B}$ and $\mathbf{R}$ are each symmetric, the two terms in Eq. (15) are transposes of one another. It is easily shown that the vanishing of their sum, for arbitrary $d\mathbf{C}$, implies the vanishing of the factors in brackets in Eq. (15). Therefore $\mathbf{C}=(1/\delta)\mathbf{B}\mathbf{R}^{-1}$, as in the 1D case.

2.2 Synchronization vs. conventional data assimilation in the linear realm

Turning now to the standard methods, so that a comparison can be made, it is recalled that the analysis $\mathbf{x}_A$ after each cycle is given by:

$$\begin{aligned} \mathbf{x}_A&=\mathbf{R}(\mathbf{R}+\mathbf{B})^{-1}\mathbf{x}_B+\mathbf{B}(\mathbf{R}+\mathbf{B})^{-1}\mathbf{x}_{\text{obs}}\\ &=\mathbf{x}_B+\mathbf{B}(\mathbf{R}+\mathbf{B})^{-1}(\mathbf{x}_{\text{obs}}-\mathbf{x}_B) \end{aligned} \tag{16}$$

In 3dVar, the background error covariance matrix $\mathbf{B}$ is fixed; in Kalman filtering it is updated after each cycle using the linearized dynamics. The background for the next cycle is computed from the previous analysis by integrating the dynamical equations:

$$\mathbf{x}_B^{n+1} = \mathbf{x}_A^n + \tau f(\mathbf{x}_A^n) \tag{17}$$

where τ is the time interval between successive analyses. Thus the forecasts satisfy a difference equation:

$$\mathbf{x}_B^{n+1} = \mathbf{x}_B^n + \mathbf{B}(\mathbf{R}+\mathbf{B})^{-1}(\mathbf{x}_{\rm obs}^n - \mathbf{x}_B^n) + \tau f(\mathbf{x}_A^n) \tag{18}$$

We model the discrete process as a continuous process in which analysis and forecast are the same:

$$\begin{aligned}\dot{\mathbf{x}}_B = f(\mathbf{x}_B) + 1/\tau\mathbf{B}(\mathbf{B}+\mathbf{R})^{-1}(\mathbf{x}_T - \mathbf{x}_B + \boldsymbol{\xi}) \\ + O[(\mathbf{B}(\mathbf{B}+\mathbf{R})^{-1})^2]\end{aligned} \tag{19}$$

using the white noise $\boldsymbol{\xi}$ to represent the difference between observation $\mathbf{x}_{\rm obs}$ and truth $\mathbf{x}_T$. The continuous approximation is valid so long as f varies slowly on the time-scale τ.

It is seen that when background error is small compared to observation error, the higher order terms $O[(\mathbf{B}(\mathbf{B}+\mathbf{R})^{-1})^2]$ can be neglected and the optimal coupling $\mathbf{C}=1/\delta\mathbf{B}\mathbf{R}^{-1}$ is just the form that appears in the continuous data assimilation equation (19), for $\delta=\tau$. Thus under the linear assumption that $Df(\mathbf{x}_B)=Df(\mathbf{x}_T)$, the synchronization approach is equivalent to 3dVar in the case of constant background error, and to Kalman filtering if background error is dynamically updated over time. The equivalence can also be shown for an exact description of the discrete analysis cycle, by comparing it to a coupled pair of synchronized maps. See Duane et al. (2006) Appendix B.

2.3 Universal covariance inflation factors in the nonlinear realm

In the fully nonlinear case, the optimal coupling scheme for synchronization may differ from that used in standard data assimilation methods.

In a region of state space where nonlinearities are strong and Eq. (9) fails, the prognostic equation for error (10) must be extended to incorporate nonlinearities. Additionally, model error due to processes on small scales that escape the digital representation should be considered. While errors in the parameters or the equations for the explicit degrees of freedom require deterministic corrections, the unresolved scales, assumed dynamically independent, can only be represented stochastically. The physical system is governed by:

$$\dot{\mathbf{x}_T} = f(\mathbf{x}_T) - \boldsymbol{\xi}_M \tag{20}$$

in place of Eq. (8a), where $\boldsymbol{\xi}_M$ is model error, with covariance $\mathbf{Q}\equiv<\boldsymbol{\xi}_M\boldsymbol{\xi}_M^T>$. The error equation (10) becomes

$$\dot{\mathbf{e}} = (\mathbf{F}-\mathbf{C})\mathbf{e} + G\mathbf{e}^2 + H\mathbf{e}^3 + \mathbf{C}\boldsymbol{\xi} + \boldsymbol{\xi}_M \tag{21}$$

where we have included terms up to cubic order in $\mathbf{e}$, with $H<0$ to prevent divergent error growth for large $||\mathbf{e}||$. In the multi-dimensional case, Eq. (21) is shorthand for a tensor equation in which G and H are tensors of rank three and rank four (and the restrictions on H are more complex). In the one-dimensional case, which we shall analyze here, G and H are scalars.

The Fokker-Planck equation is now:

$$\frac{\partial \rho}{\partial t} + \nabla_e \cdot \{\rho[(F - C)e + Ge^2 + He^3]\} = \frac{1}{2}\delta \nabla_e \cdot [(CRC^T + Q)\nabla_e \rho] \tag{22}$$

Using the ansatz for the PDF ρ:

$$\rho(e) = N \exp(-Ke^2 - Le^3 - Me^4) \tag{23}$$

with a normalization factor $N = [\int_{-\infty}^{\infty} de \ \exp(-Ke^2 - Le^3 - Me^4)]^{-1}$, we obtain from Eq. (22) the following relations between the dynamical parameters and the PDF parameters:

$$\begin{aligned} F - C &= \frac{1}{2}\tau(C^2R + Q)(-2K) \\ G &= \frac{1}{2}\tau(C^2R + Q)(-3L) \\ H &= \frac{1}{2}\tau(C^2R + Q)(-4M) \end{aligned} \tag{24}$$

The goal is to minimize the background error:

$$B(K, L, M) = \frac{\int_{-\infty}^{\infty} de \ e^2 \exp(-Ke^2 - Le^3 - Me^4)}{\int_{-\infty}^{\infty} de \ \exp(-Ke^2 - Le^3 - Me^4)}. \tag{25}$$

Using Eq. (24) to express the arguments of B in terms of the dynamical parameters, we find $B(K, L, M)=B(K(C), L(C), M(C))\equiv B(C)$, and can seek the value of C that minimizes B, for fixed dynamical parameters F, G, H.

The coupling that gives optimal synchronization can again be compared with the coupling used in standard data assimilation, as for the linear case. In particular, one can ask whether the "covariance inflation" scheme that is used as an ad hoc adjustment in Kalman filtering (Anderson 2001) can reproduce the C values found to be optimal for synchronization. The form $C=\tau^{-1}B(B+R)^{-1}$ is replaced by the adjusted form

$$C = \frac{1}{\tau}\frac{\mathcal{F}B}{\mathcal{F}B + R} \tag{26}$$

where $\mathcal{F}$ is the covariance inflation factor.

The optimization problem was solved numerically in the one-dimensional case (Duane et al. 2006) with results as plotted in Table 17.1. If the function f in the dynamical equation describes motion in a two-well potential, with minima at distances d_1 and d_2 from the unstable fixed point, it can be shown that the dynamical parameters in (21) are $G=1/d_2-1/d_1$ and $H=-1/(d_1d_2)$. Results are displayed in the table for

Table 17.1. Covariance inflation factor vs. bimodality parameters d_1, d_2, for 50% model error in the resolved tendency.

	d_1					
	.75	1.	1.25	1.5	1.75	2.
.75	1.26	1.26	1.28	1.30	1.32	1.34
1.	1.26	1.23	1.23	1.25	1.27	1.29
d_2 1.25	1.28	1.23	1.22	1.23	1.24	1.25
1.5	1.30	1.25	1.23	1.22	1.23	1.24
1.75	1.32	1.27	1.24	1.23	1.23	1.23
2.	1.34	1.29	1.25	1.24	1.23	1.23

a range of values of the bimodality parameters d_1 and d_2. The model error in Eq. (20) is chosen to be about 50% of the resolved tendency $\dot{x}_T$, with the resulting model error covariance Q=0.04 approximately one-fourth of the background error covariance B. The covariance inflation factors are remarkably constant over a wide range of parameters and agree with typical values used in operational practice.

3 Automatic parameter estimation and model learning in synchronization-based data assimilation

3.1 Identical synchronization implies parameter estimation for non-identical systems

Machine learning might also be realized in the synchronization context, so as to correct for deterministic model error in the resolved degrees of freedom. By allowing model parameters to vary slowly, generalized synchronization would be transformed to more nearly identical synchronization. Indeed, parameter adaptation laws can be added to a synchronously coupled pair of systems so as to synchronize the parameters as well as the states. Parlitz (1996) showed for example that two unidirectionally coupled Lorenz systems with different parameters:

$$\begin{aligned} \dot{X} &= \sigma(Y - X) & \dot{X}_1 &= \sigma(Y - X_1) \\ \dot{Y} &= \rho X - Y - XZ & \dot{Y}_1 &= \rho_1 X_1 - \nu Y_1 - X_1 Z_1 + \mu \\ \dot{Z} &= -\beta Z + XY & \dot{Z}_1 &= -\beta Z_1 + X_1 Y_1 \end{aligned} \tag{27}$$

could be augmented with parameter adaptation rules:

$$\begin{aligned} \dot{\rho}_1 &= (Y - Y_1)X_1 \\ \dot{\nu} &= (Y_1 - Y)Y_1 \\ \dot{\mu} &= Y - Y_1 \end{aligned} \tag{28}$$

so that the Lorenz systems would synchronize, and additionally $\rho_1 \to \rho$, $\nu \to 1$, and $\mu \to 0$.

Equations for a synchronously coupled pair of systems can in fact always be augmented to allow parameter adaptation as well, provided that relevant dynamical variables are observed, as shown by Duane, Yu, and Kocarev (2007). Their findings are reproduced in this section.

Consider a "real system" given by ODE's:

$$\dot{\mathbf{x}} = \boldsymbol{f}(\mathbf{x}, \mathbf{p}), \tag{29}$$

$$\dot{\mathbf{p}} = 0, \tag{30}$$

where $\mathbf{x} \in I\!R^N$, $\boldsymbol{f} : I\!R^N \rightarrow I\!R^N$, and $\mathbf{p} \in I\!R^m$ is the vector of (unknown, constant) parameters of the system. We further assume that $\mathbf{s} = \mathbf{s}(\mathbf{x})$, where $\mathbf{s} : I\!R^N \rightarrow I\!R^n$, $n \leq N$, is an n dimensional vector representing the experimental measurement output of the system. A "computational model" of the system (29) is given by:

$$\dot{\mathbf{y}} = \boldsymbol{f}(\mathbf{y}, \mathbf{q}) + \boldsymbol{u}(\mathbf{y}, \mathbf{s}), \tag{31}$$

$$\dot{\mathbf{q}} = \boldsymbol{N}(\mathbf{y}, \mathbf{x} - \mathbf{y}), \tag{32}$$

where $\boldsymbol{N}(\mathbf{y}, 0) = 0$, and $\boldsymbol{u}$ is the control signal. Let $\mathbf{e} \equiv \mathbf{y} - \mathbf{x}$ and $\mathbf{r} \equiv \mathbf{q} - \mathbf{p}$. Choose a positive definite Lyapunov function $L_o(\mathbf{e})|_{\mathbf{q}=\mathbf{p}}$. Assume that the control signal $\boldsymbol{u}$ is designed such that there is some time t_0 for which $\dot{L}_o(\mathbf{e}(t))|_{\mathbf{q}=\mathbf{p}} < 0$ when $\mathbf{e}(t) \neq 0$ and $\dot{L}_o(\mathbf{e}(t))|_{\mathbf{q}=\mathbf{p}} = 0$ when $\mathbf{e}(t) = 0$, for all $t > t_0$. That is, after time t_0, the system proceeds monotonically toward synchronization. Let $\boldsymbol{h} \equiv \boldsymbol{f}(\mathbf{y}, \mathbf{r} + \mathbf{p}) - \boldsymbol{f}(\mathbf{y} - \mathbf{e}, \mathbf{p})$. Duane et al. (2007) established the following theorem:

Theorem 1. *Assume that (i) the control law* $\boldsymbol{u}$ *in (31) is designed such that the synchronization manifold* $\mathbf{x} = \mathbf{y}$ *is globally asymptotically stable, (ii)* $\boldsymbol{f}$ *is linear in the parameters* $\mathbf{p}$*, and (iii) the parameter estimation law (32) is designed such that*

$$N_j = -\delta_j \sum_i \left(\frac{\partial L_o}{\partial e_i} \right) \left(\frac{\partial h_i}{\partial r_j} \right),$$

where δ_j *are positive constants. Then the synchronization manifold* $\mathbf{y} = \mathbf{x}$, $\mathbf{p} = \mathbf{q}$ *is globally asymptotically stable.*

The theorem ensures the stability of the synchronization manifold $\mathbf{y} = \mathbf{x}, \mathbf{p} = \mathbf{q}$. It says that if the two systems synchronize for the case of identical parameters, then the parameters of the "real system" can be estimated when they are not known *a priori*, provided that each partial derivative $\partial L_o / \partial e_i$ is known for which the vector $\partial h_i / \partial r_j$ $(j = 1, ...)$ is not zero. For the usual form $L_o \equiv \sum_i (e_i)^2$, the requirement is that x_i be known if the equation for $\dot{y}_i$ contains parameters that one seeks to estimate. By considering a more general Lyapunov function that is defined in terms of some subset S of the state variables, or their indices, $L_o \equiv \sum_{i \in S} c_i (e_i)^2$ for positive coefficients c_i, one obtains the looser requirement for each desired parameter, that x_i be known for at least some i for which the $\dot{y}_i$ equation contains that parameter. (Convergence may be slower if fewer x_i are known.)

3.2 Parameter estimation in geophysical fluid models

Generalizations to PDEs would allow parameters in geophysical models to automatically adapt. In general, some of the partial derivatives $\partial L_o/\partial e_i$ may not be known and Theorem 1 may be inapplicable. The theorem gives us no way to estimate the parameters ρ or b for the Lorenz system, for instance, if only x_1 is coupled, while x_2 and x_3 (hence $\partial L_o/\partial e_2$ and $\partial L_o/\partial e_3$) are unknown. But in the case of translationally invariant PDE's, the parameters are the same at each point in space. In general, they can be estimated from a limited amount of information about the state at a discrete set of points (or for a finite set of Fourier components), if such information is also sufficient to give identical synchronization globally or locally when coupled to a "model system". Consider a pair of spatially extended systems that are synchronously coupled in one direction:

$$\begin{aligned} \frac{\partial \phi^A}{\partial t} + \Gamma(\phi^A) &= f^A F(\phi^A, \phi^A) \\ \frac{\partial \phi^B}{\partial t} + \Gamma(\phi^B) + C(\phi^A, \phi^B) &= f^B F(\phi^B, \phi^A), \end{aligned} \tag{33}$$

where Γ is some general form in the field ϕ, possibly involving spatial derivatives, C is a similarly general form coupling ϕ^B to ϕ^A, such that $C(\phi, \phi) = 0$, and the form F with coefficient f specifies a "forcing", which in general may also couple to the other system. The two systems are dynamically identical when $\phi^A = \phi^B$ and $f^A = f^B$.

Define a core Lyapunov function $L_o(\phi^A - \phi^B)|_{f^A = f^B} \equiv \int d^3x\ (\phi^A - \phi^B)^2$. Then a ready generalization of Theorem 1, to the case of a continuum of state variables tells us that if we impose the parameter estimation law

$$\dot{f}^B = \int d^3x\ (\phi^A - \phi^B) F(\phi^B, \phi^A), \tag{34}$$

we will find $f^B \to f^A$ as in the ODE case.

The procedure outlined above estimates a coefficient of forcing for a wide class of synchronously coupled PDE's, and is readily generalized for limited measurements of ϕ^A. The quasigeostrophic potential vorticity equation, for instance, used to describe the large-scale atmospheric circulation, has been shown to exhibit high-quality synchronization when two copies are coupled via partial exchange of only mid-range Fourier components of the flow field (Duane and Tribbia 2001, Duane and Tribbia 2004). The synchronization manifold is either globally attracting or locally attracting in a very wide basin. Synchronization of such systems may be useful for meteorological data assimilation (Duane 2003). It is important that forcing parameters in the equation can also be estimated.

Consider a two-layer channel model (derived from one described by Vautard et al. (1988)), with flow evolution given by the quasigeostrophic equation for potential vorticity q on a β-plane:

$$\frac{Dq_l}{Dt} \equiv \frac{\partial q_l}{\partial t} + J(\psi_l, q_l) = F_l + D_l \tag{35}$$

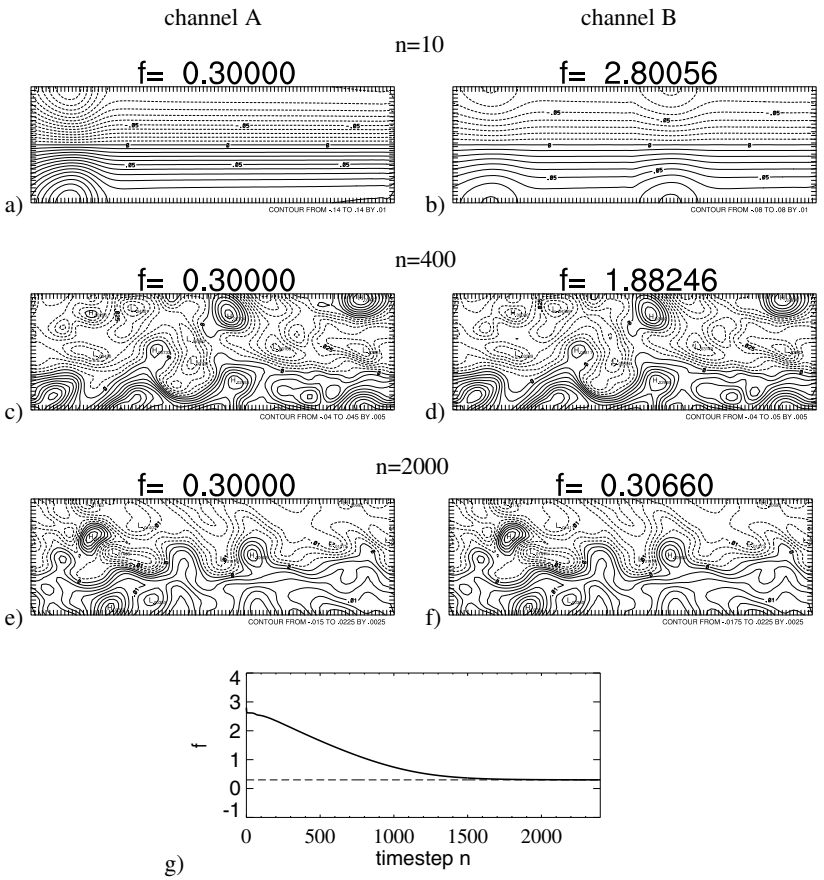


Fig. 17.4. The evolving flow ψ (a-f) for two quasigeostrophic channel models that are synchronously coupled as in (Duane and Tribbia 2001; Duane and Tribbia 2004) (but in one direction only), and with the forcing parameter f^B for the second channel (denoted μ_o in the reference) allowed to vary according to the truncated parameter adaptation rule (37) with $S = \{\mathbf{k} : \mathbf{k}_x, \mathbf{k}_y \leq 12\}$ (in waves per channel- length). Starting from the initial value $f^B = 3.0$ at time step $n = 0$ (not shown), f^B converges (g) to the value of the corresponding parameter $f^A = 0.3$ (dashed line) in the first channel, as the flows synchronize. (An average of the two layers $l = 1, 2$ is shown.)

where the layer l=1, 2, ψ is streamfunction, and the Jacobian $J(\psi, \cdot) = \frac{\partial \psi}{\partial x}\frac{\partial \cdot}{\partial y} - \frac{\partial \psi}{\partial y}\frac{\partial \cdot}{\partial x}$ gives the advective contribution to the Lagrangian derivative D/Dt. Equation (35) states that potential vorticity is conserved on a moving parcel, except for forcing F_l and dissipation D_l. The discretized potential vorticity is $q_l = f_0 + \beta y + \nabla^2 \psi_l + R_l^{-2}(\psi_1 - \psi_2)(-1)^i$ where $f(x, y)$ is the vorticity due to the Earth's rotation at each point (x, y), f_0 is the average f in the channel, β is the constant df/dy and R_l is the Rossby radius of deformation in each layer. The forcing F is a relaxation term designed to induce a jet-like flow near the beginning of the channel: $F_l = f(q_l^* - q_l)$ for q_l^* corresponding to a choice of ψ^* that resembles the flow in Fig. 17.4a, for example.

The dissipation terms D, boundary conditions, and other parameter values are given in Duane and Tribbia (2004). One might seek to estimate the coefficient f.

Consider two systems, "A" and "B", both given by equations of the form (35), but with the forcing in the B system defined differently, in terms of its spectral components $F^B_{\mathbf{k}}$, so as to effect a unidirectional coupling:

$$F^B_{\mathbf{k}} = f^B \sum_{\mathbf{k}} a_{\mathbf{k}}(q^*_{\mathbf{k}} - q^B_{\mathbf{k}}) + f^B \sum_{\mathbf{k}} b_{\mathbf{k}}(q^A_{\mathbf{k}} - q^B_{\mathbf{k}}) \tag{36}$$

where the layer index l is suppressed and the coefficients $a_{\mathbf{k}}, b_{\mathbf{k}}$ are slightly smoothed step functions of $\mathbf{k}$, so that each spectral component is either coupled to the corresponding component in the A system or to the background flow q^* or niether. The coefficients are chosen so as to couple only the medium-scale components:

$$b_{\mathbf{k}} = \begin{cases} 0 & \text{if } |k_x| \le k_{x0} \text{ and } |k_y| \le k_{y0} \\ (k_n/|\mathbf{k}|)^4 & \text{if } |\mathbf{k}| > k_n \\ 1 - (k_0/|\mathbf{k}|)^4 & \text{otherwise} \end{cases}$$

and to force only the large-scale components:

$$a_{\mathbf{k}} = \begin{cases} 1 - b_{\mathbf{k}} & \text{if } |\mathbf{k}| \le k_n \\ 0 & \text{if } |\mathbf{k}| > k_n \end{cases}$$

as in (Duane and Tribbia 2004), where the constants k_0, k_{x0}, k_{y0} and k_n are defined. The systems thus coupled synchronize, without bursting, as shown in Fig. 17.4. (The forcing for the A system is correspondingly truncated: $F^A_{\mathbf{k}} = f^A \sum_{\mathbf{k}} a_{\mathbf{k}}(q^*_{\mathbf{k}} - q^A_{\mathbf{k}})$.)

The parameter estimation rule in spectral space:

$$\dot{f}^B = \sum_{\mathbf{k}\in S} (q^A_{\mathbf{k}} - q^B_{\mathbf{k}})[a_{\mathbf{k}}(q^*_{\mathbf{k}} - q^B_{\mathbf{k}}) + b_{\mathbf{k}}(q^A_{\mathbf{k}} - q^B_{\mathbf{k}})] \tag{37}$$

is the Fourier transform of (34) if S is universal. But even for a restricted range of wavenumbers in S, as in the figure, the rule (37) causes f^B to converge to f^A as would follow from the use of a correspondingly restricted Lyapunov function.

As a more realistic example than the channel model described above, the Weather Research and Forecasting (WRF) model was considered, as adapted for weather prediction over military test ranges for the Army Test and Evaluation Command (ATEC). The ATEC application is based on observations that are so frequent that they can be assumed to occur at every numerical time step, so that a continuously coupled differential equation system of the form (8) or (19) can be taken to reflect the actual data assimilation scenario.

At a relevant level of model detail, the prognostic equation for humidity (water vapor mixing ratio) q is:

$$\frac{\partial q}{\partial t} = \frac{\partial}{\partial z}\{K(\frac{\partial q}{\partial z} - Mf(u_0, T_0, ...))\} \tag{38}$$

where K is a moisture diffusivity, and $M = M(x, y)$ quantifies the impact of soil moisture at each location (x, y), which is a function f of state variables u_0, T_0, etc. at the surface. To study the estimation of M using the synchronization method, attention is restricted to a single vertical column $(x, y) = (x_0, y_0)$ and a model is introduced that is diffusively coupled to ("hudged"by) the true state. The model humidity q_m, for instance, is governed by:

$$\frac{\partial q_m}{\partial t} = \frac{\partial}{\partial z}\{K(\frac{\partial q_m}{\partial z} - Mf(u_{m0}, T_{m0}, ...))\} + c(q_{obs} - q_m) \tag{39}$$

where q_{obs} is the observed humidity (at any level z where an observation is taken) that is the sum of the true q and observational noise. c is a coupling ("hudging") coefficient. Similar equations govern the evolution of temperature T, wind speed u, and other model variables, but the parameter M is thought to enter only the humidity equation (39).

In accordance with the general rule (34), M for the model was made to vary with observational input as:

$$\dot{M} \sim -\frac{\partial K}{\partial z} f(u_{m0}, T_{m0}, ...)(q_{obs}(z) - q_m(z)) \tag{40}$$

for the case of observations taken at just one level z.

Repeated convergence of M to its true value, each time followed by a burst away from synchronization, is seen in Fig. 17.5.

3.3 Stochasticity for global optimization and model learning: a heuristic example

In complex cases, where there are multiple zeros of $\dot{L}$ corresponding to local optima, a stochastic component (in the parameters) may be necessary to allow parameters to jump among multiple basins of attraction so as to reach the global optimum (Fig. 17.6).

To illustrate the utility of added noise in optimizing synchronization patterns, we turn to a synchronized oscillator ("cellular neural network (CNN)") representation of the travelling salesman problem. The representation generalizes an older one due to Hopfield and Tank (1985) a traditional neural network, fully interconnected, with fixed weights chosen so that the globally optimal state corresponds to a shortest-distance "tour"among a collection of cities with a pre-specified distance for each pair of cities. In this original representation, for a collection of 5 cities, one considers a 5×5 matrix of binary values, where the rows correspond to cities and the columns correspond to slots, 1 through 5, in the tour schedule. A tour is any pattern of 0's and 1's, such that there is exacly one 1 in each row (exactly one city visited at a time) and one 1 in each column (each city visited exactly once). For instance the tour depicted in Fig. 17.7a is ECABD. There is a 10-fold degeneracy in optimal patterns (shortest-distance cyclic tours) due to arbitrariness in the selection of the starting city and the direction of the tour. The travelling salesman problem, in this representation or any other, is difficult because of the multiplicity of local optima.

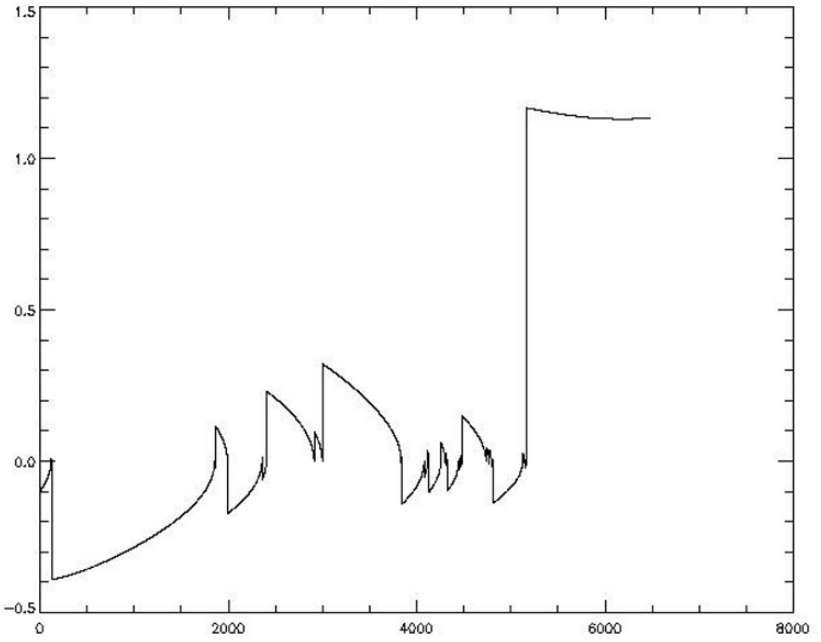


Fig. 17.5. The variable model parameter M converges to the true value M_T with repeated "bursting", in the dynamical parameter adaptation scheme that only requires a single realization ($M - M_T$ is plotted). Results are unstable, but even for this case, where observations are only assimilated at one point, the correct value of the parameter can be identified. (State variables also do not converge completely over the time interval shown.)

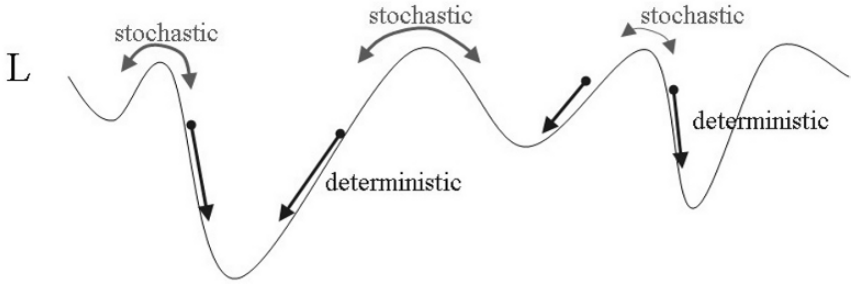


Fig. 17.6. Deterministic parameter estimation rules cause parameters to reach local optima of the Lyapunov function L. Stochasticity (e.g. in a simulated annealing algorithm) allows jumps among different basins of attraction.

The representation considered here is a 5×5 array of coupled periodic oscillators. A tour is specified by a synchronization pattern in which all oscillators in each row and each column are desynchronized, and such that for each oscillator in each column (row), there is exactly one oscillator in each other column (row) that is synchronized with it. In Fig. 17.7b, the same tour ECABD is depicted in the new representation, simultaneously

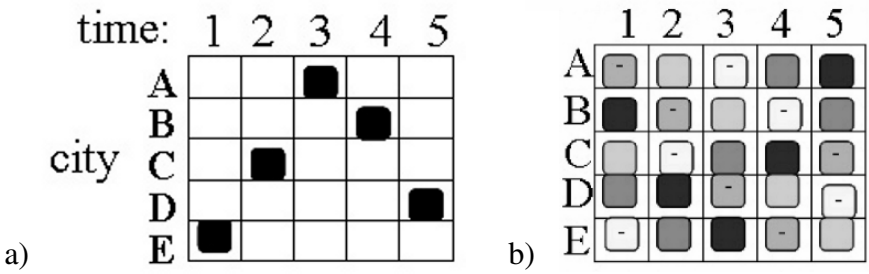


Fig. 17.7. In the original Hopfield representation (a), a tour (here ECABD) is specified by the collection of units that are "on", provided there is only one "on" unit in each row and only one in each column. In the CNN representation (b) equivalent cyclic permutations of the same tour are simultaneously represented as sets of 5 units that are synchronized, with an analogous proviso. Units with the same relative phase are shown in the same color. (Dash marks distinguish ambiguous colors in the grayscale version.)

with equivalant tours given by cyclic permutations of cities: CABDE, ABDEC, etc. There is now only a two-fold degeneracy in optimal patterns, due to arbitrariness in direction.

To solve the travelling salesman problem in the representation we have described, let each oscillator be given by a complex number z_{ij} $(i = 1, \ldots 5, j = A, \ldots E)$ that contains both a phase $\arg(z_{ij})$ and an amplitude $|z_{ij}|$. Assume all oscillators have the same frequency ω and make the replacement $z_{ij} \rightarrow \exp(-i\omega t) z_{ij}$, so that only the relative phases are represented in the complex quantities z_{ij}. The Lyapunov function we seek to minimize is:

$$L = A \sum_{ij} \left| |z_{ij}|^2 - 1 \right|^2 + B \sum_{ij} \left| \left(\frac{z_{ij}}{|z_{ij}|} \right)^5 - 1 \right|^2 - C \sum_{i} \sum_{jj'} \left| \frac{z_{ij}}{|z_{ij}|} - \frac{z_{ij'}}{|z_{ij'}|} \right|^2$$
$$-D \sum_{j} \sum_{ii'} \left| \frac{z_{ij}}{|z_{ij}|} - \frac{z_{i'j}}{|z_{i'j}|} \right|^2 + E \sum_{i} \sum_{j} \sum_{j'} d_{jj'} \mathrm{Re}\left(\frac{z_{ij}}{|z_{ij}|} \frac{z^*_{i+1,j'}}{|z_{i+1,j'}|} \right) \qquad (41)$$

where asterisks denote complex conjugates, and we close the tours by defining $z_{6j} \equiv z_{1j}$. The first term, with coefficient A, tends to force $|z| = 1$. The second term, with coefficient B tends to force z to one of five phase states, corresponding to the fifth roots of unity. The terms with coefficients C and D penalize for synchronization within each row and within each column, respectively. The last term, expressed in terms of distances $d_{jj'}$ between cities j and j', is easily seen to be minimized when each partial sum $\sum^{sync}$ over synchronized oscillators at i, j and $i+1, j'$ (i.e. for which $\arg(z_{ij}) = \arg(z_{i+1,j'})$), $\sum_{ijj'}^{sync} d_{jj'}$ is minimized, that is when the specified tour has shortest distance.

If the oscillators are governed by equations

$$\dot{z}_{ij} = -\frac{\partial L}{\partial z^*_{ij}}, \qquad \dot{z}^*_{ij} = -\frac{\partial L}{\partial z_{ij}} \qquad (42)$$

Table 17.2. Relative phases (from $-\pi$ to π) of oscillators in the 5×5 CNN representation of the 5-city travelling salesman problem, after reaching a steady state. No tours are evident.

	slot in schedule				
	1	2	3	4	5
A	-2.47002	-2.18725	-0.14433	1.28413	1.28601
B	2.49589	-1.22075	1.40448	-2.48385	-0.01186
city C	1.23951	2.64314	-2.61880	0.00870	-1.22135
D	0.00406	1.08844	-2.63690	-1.27138	2.53321
E	-1.29235	1.11239	0.04947	2.51944	-2.52968

Table 17.3. Relative phases as in Table 17.2, but with noise of steadily decreasing amplitude included in the oscillator dynamics (43). The tour found, ADEBC, is seen by focussing attention on oscillators with relative phase $-2\pi/5 \approx -1.28$. Permutations of the same tour are represented by other collections of oscillators with approximately equal relative phases. ADEBC is indeed the shortest-distance tour for the pre-specifi ed table of distances ($d_{jj'}$) that was arbitrarily selected.

	slot in schedule				
	1	2	3	4	5
A	-1.28206	0.03570	1.23972	2.49383	-2.52120
B	1.23280	2.48519	-2.49996	-1.31980	-0.01645
city C	0.03447	1.27973	2.52424	-2.53673	-1.25935
D	-2.49808	-1.29343	-0.01308	1.19502	2.50160
E	2.49330	-2.42328	-1.19018	0.04572	1.28507

following Hoppensteadt (1996), then the derivative of the Lyapunov function $\dot{L} = \sum_{ij}[(\partial L/\partial z_{ij})\,\dot{z}_{ij} + (\partial L/\partial z^*_{ij})\,\dot{z}^*_{ij}] = -2\sum_{ij}|\partial L/\partial z_{ij}|^2 \leq 0$, so that the cost function is monotonically decreasing and must reach at least a local minimum, since L can be easily seen to be bounded below.

The point of this lengthy description is that the network as described almost never reaches a global minimum, or even a state defining a tour. For a given set of prespecified distances (not shown) the network reaches a phase state shown in Table 17.2. However, if gaussian noise is added to the dynamics

$$\dot{z}_{ij} = -\frac{\partial L}{\partial z^*_{ij}} + \xi, \qquad \dot{z}^*_{ij} = -\frac{\partial L}{\partial z_{ij}} + \xi^* \tag{43}$$

where ξ represents complex gaussian noise, then the system attains the tour state shown in Table 17.3, which furthermore is the shortest-distance tour, provided that the amplitude of the noise (an analog "temperature") is decreased slowly enough. This technique of "simulated annealing", which had also been applied in the original Hopfield representation, is effective in selecting one of a small number of global optima (precisely $2 \times 5! = 240$) out of 5^{25} phase states. (The term in (41) restricting the phase of each oscillator to one of five values, can be dropped, but the the temperature must be lowered even more slowly to attain a globally optimal pattern.)

The complexity of this combinatorial example far exceeds that of most parameter estimation problems, in which there are usually only a small number of local optima.

But the example suggests that simulated annealing could perhaps be extended to a genetic algorithm that would make random qualitative changes in the model as well, until synchronization is achieved. The approach to model learning thus defined is in keeping with the suggestion that synchronicity is a useful paradigm for the relationship between reality and a computational model.

Acknowledgments

The authors thank Jeff Weiss, Dongchuan Yu, Ljupco Kocarev, and Josh Hacker for useful discussions. This work was partially supported under NSF Grant 0327929. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

References

Afraimovich, V. S., Verichev, N. N., and Rabinovich, M. I. (1986) Stochastic synchronization of oscillation in dissipative systems. Radiophys. Quantum Electron. 29, 795-803.

Anderson, J. L. (2001) An ensemble adjustment Kalman filter for data assimilation. Mon. Wea. Rev. 129, 2884-2903.

Anderson, J. L. (2003) A local least-squares framework for ensemble filtering. Mon. Wea. Rev. 131, 634-642.

Daley, R. (1991) Atmospheric Data Analysis, Cambridge Univ. Press, Cambridge.

Duane, G.S. (1997) Synchronized chaos in extended systems and meteorological teleconnections. Phys. Rev. E 56, 6475-6493.

Duane, G.S. (2003) Synchronized chaos in climate dynamics, in: In, V., Kocarev, L., Carroll, T.L. et al. (eds.), AIP Conference Proceedings 676, Melville, New York, pp. 115-126.

Duane, G.S. and Tribbia, J.J. (2001) Synchronized chaos in geophysical fluid dynamics. Phys. Rev. Lett. 86, 4298-4301.

Duane, G.S. and Tribbia, J.J. (2004) Weak Atlantic-Pacific teleconnections as synchronized chaos. J. Atmos. Sci. 61, 2149-2168.

Duane, G.S., Tribbia, J.J., and Weiss, J.B. (2006) Synchronicity in predictive modelling: a new view of data assimilation. Nonlin. Processes in Geophys. 13, 601-612.

Duane, G.S., Yu, D., and Kocarev, L. (2007) Identical synchronization, with translation invariance, implies parameter estimation. Phys. Lett. A. (in press)

Fujisaka, H. and Yamada, T. (1983) Stability theory of synchronized motion in coupled-oscillator systems. Prog. Theor. Phys. 69, 32-47.

Hopfield, J.J. and Tank, D.W. (1985)"Neural" computation of decisions in optimization problems. Biol. Cybern. 52, 141-152.

Hoppensteadt, F.C. (1996) Synaptic organizations and dynamical properties of weakly connected neural oscillators 2. Learning phase information. Biol. Cybern. 75, 129-135.

Huygens, C. (1673) *Horoloqim Oscillatorium.* Apud. F. Muget, Paris.

Jung, C. G. and Pauli, W. (1955) The interpretation of nature and the psyche, Pantheon, New York.

Kocarev, L., Tasev, Z., and Parlitz, U. (1997) Synchronizing spatiotemporal chaos of partial differential equations. Phys. Rev. Lett. 79, 51-54.

Lorenz, E. N. (1963) Deterministic nonperiodic flows. J. Atmos. Sci. 20, 130-141.

Parlitz, U. (1996) Estimating model parameters from time series by autosynchronization. Phys. Rev. Lett. 76, 1232-1235.

Parlitz, U. and Kocarev, L. (1997) Using surrogate data analysis for unmasking chaotic communication systems. Int. J. Bifurcations and Chaos 7, 407-413.

Peat, F. D. (1987) Synchronicity: The Bridge Between Matter and Mind, Bantam, New York.

Pecora, L. M. and Carroll, T. L. (1990) Synchronization in chaotic systems. Phys. Rev. Lett. 64, 821-824.

Pecora, L. M., Carroll, T. L., Johnson, G. A., Mar, D. J., and Heagy, J. F. (1997) Fundamentals of synchronization in chaotic systems, concepts, and applications. Chaos 7, 520-543.

Rulkov, N. F., Sushchik, M. M., and Tsimring, L. S. (1995) Generalized synchronization of chaos in directionally coupled chaotic systems. Phys. Rev. E 51, 980-994.

So, P., Ott, E., and Dayawansa, W. P. (1994) Observing chaos – deducing and tracking the state of a chaotic system from limited observation. Phys. Rev. E 49, 2650-2660.

Strogatz, S. H. (2003) Sync: The Emerging Science of Spontaneous Order. Theia, New York.

Vautard, R., Legras, B., and Déqué, M. (1988) On the source of midlatitude low-frequency variability. Part I: A statistical approach to persistence. J. Atmos. Sci. 45, 2811-2843.

Von Der Malsburg, C. and Schneider, W. (1986) A neural cocktail-party processor. Biol. Cybernetics 54, 29-40.

Yang, S.-C., Baker, D., Cordes, K., Huff, M., Nagpal, G., Okereke, E., Villafañe, J., and Duane, G. S. (2006) Data assimilation as synchronization of truth and model: Experiments with the three-variable Lorenz system. J. Atmos. Sci. 63, 2340-2354.

18 Scale, Scaling and Multifractals in Geophysics: Twenty Years on

Shaun Lovejoy[1] and Daniel Schertzer[2,3]

[1] Physics, McGill University, 3600 University st., Montreal, Que. H3A 2T8, Canada, lovejoy@physics.mcgill.ca
[2] CEREVE, Ecole Nationale des Ponts et Chaussées,6-8, avenue Blaise Pascal, Cité Descartes, 77455 MARNE-LA-VALLEE Cedex, France, Daniel.Schertzer@enpc.fr
[3] CNRM, Météo-France, 1 Quai Branly, 75007 Paris, France

Abstract. We consider three developments in high number of degrees of freedom approaches to nonlinear geophysics: a) the transition from fractal geometry to multifractal processes, b) the self-organized critical (SOC) generation of extremes via multifractal phase transitions c) the generalization from isotropic scale invariance (self-similar fractals, multifractals) to (anisotropic) generalized scale invariance. We argue that these innovations are generally necessary for geophysical applications. We illustrate these ideas with data analyses from both the atmosphere and the earth's surface, as well as with multifractal simulations.

1 Introduction: Which Chaos in Geophysics?

During the 1960's, 70's and 80's theoretical developments in geophysics, physics, and mathematics spawned four related non-linear paradigms. The first, deterministic chaos, was centred around the discovery [*Lorenz*, 1963], [*Ruelle and Takens*, 1971] that systems with as few as three degrees of freedom could have random-like "chaotic" behaviour. The second, fractal geometry - proposed that many natural systems could be modeled as (stochastic, scale invariant) fractal sets [*Mandelbrot*, 1967]. The third, "self-organized criticality" (SOC) [*Bak, et al.*, 1987] proposed that extreme events could be the result of seemingly simple generic avalanche-like processes. As discussed below SOC turns out to have close relationships with *non classical* critical phase transitions, which help to establish links to the fourth scaling tool, the "renormalization group". The latter has been extremely helpful in clarifying classical phase transitions [*Wilson*, 1971] and in understanding the dynamical nature of scaling. By the early 1980's further developments had made the first two quite practical. In particular, the discovery of universality in chaos by [*Feigenbaum*, 1978] had emphasized the generic features of chaotic models rendering them more applicable, and the revolution in computer graphics had made fractals – including "strange" chaotic attractors – palpable.

This short expose gives an overview of certain subsequent developments covering roughly the twenty years celebrated by the conference. While deterministic chaos is

essentially a low degrees of freedom paradigm, fractals and SOC are both high number of degrees of freedom frameworks and could thus be called "stochastic chaos" since they involve infinite dimensional probability spaces [*Lovejoy and Schertzer*, 1998a]. While the question of whether deterministic or stochastic chaos is more geophysically relevant continues to be debated, here we focus on the latter (in particular *Schertzer et al.*, 2002 who showed that a finite correlation dimension does not discriminate between deterministic and stochastic chaos). We outline three key developments which allow the fractal and SOC paradigms to be widely applicable in geosciences: a) the transition from scale invariant sets (fractals) to scale invariant fields (multifractals), b) the recognition of the link between extreme events, heavy tailed (algebraic) probabilities (SOC) and space-time scaling, c) the generalization of scale invariance from isotropic (self-similar) systems to very general anisotropic ones (self-affine and beyond) within the framework of Generalized Scale Invariance (GSI). Rather than attempt a systematic survey, we illustrate our discussion using results from key fields in solid earth and atmospheric geophysics: the earth's topography and clouds and rain. This choice is motivated by both the fundamental significance of the fields and for the availability of relevant high quality data.

2 The Link between Descriptions and Models

It is an old truism that one cannot make a measurement without first having a theory of what is to be measured. This is well illustrated in nonlinear geophysics where theoretical developments are not only necessary for making more sophisticated theories and better applications: they are necessary simply in order to quantitatively *describe* geofields. We illustrate this statement with two significant examples. The first is the long debate starting in the 1980's about what was the (supposedly unique) fractal dimension of the earth's topography. If the topography could be adequately modeled as a geometrical fractal set, then many different techniques (including spectral analysis) could be used to estimate its unique dimension D. Unfortunately, different techniques applied to different data commonly gave different values of D (see the review in [*Klinkenberg and Goodchild*, 1992]). Consequently by the end of the 1990's the mainstream surface geomorphology community had "moved on", relegating fractals to narrow ranges of scale and to very technical applications. This near abandonment of scaling occurred in spite of the fact that entire fields of research such as surface hydrology (see e.g. the excellent review [*Rodriguez-Iturbe and Rinaldo*, 1997]) are riddled with scaling laws which virtually require the topography to respect some form of scaling. At the same time, due to their random singularities, multifractals have such strong variability that they violate many conventional geostatistical assumptions so that normal multifractal variability can easily be misinterpreted in terms of spurious scale breaks, spurious nonstationarity etc. The loss of interest in scaling was encouraged by the extensive use of (low variability) fractional Brownian motion (fBm) models of topography. As argued in [*Gagnon, et al.*, 2006], the topography in fact has excellent multiscaling (multifractal) properties (see Fig. 18.2, 18.4, 18.7) – but an infinite hierarchy of fractal dimensions; this requires new analysis techniques. Consequently

the lack of an adequate theoretical framework for scaling has led the baby to be thrown out with the bathwater.

Similarly, in the atmosphere the still dominant model of atmospheric dynamics is of a small scale 3D isotropic turbulence [*Monin*, 1972], [*Orlanski*, 1975] superposed on a large scale 2D isotropic turbulence with the two separated by a "mesoscale gap" somewhere near the 10km scale thickness (the "meso-scale gap", [*Van der Hoven*, 1957]). The empirical evidence of this energy gap in the atmospheric spectrum had been more and more put into question, in particular by (*Gage*, 1979; *Lilly and Paterson*, 1983). The drastic consequences of such a "dimensional transition" were theoretically put forward by (*Schertzer and Lovejoy*, 1984; *Schertzer and Lovejoy*, 1985), and to rather consider a unique scaling regime corresponding to a model which is neither 2D nor 3D, but with an "elliptical dimension" $D_{el} = 23/9 \approx 2.55$. The 2D/3D model is increasingly at odds with modern data – particularly of the vertical structure - which fails to find evidence of any isotropic turbulent regime whatsoever (see [*Lovejoy, et al.*, 2007])! Indeed, the mainstream experimentalists espouse anisotropic but scaling gravity wave models (e.g. [*Dewan*, 1997] [*Gardner*, 1994]) which are equivalent to "elliptical dimensions" (see below) $D_{el} = 7/3$ (in between 2 and 3), whereas high resolution vertical cross data (from lidars) favour a value closer to 23/9 [*Lilley, et al.*, 2004]. Today, probably the key element blocking a consensus is the fact that there is still no general agreement about the horizontal structure. This is where a marriage between theory and measurements is required: how to interpret the aircraft data which are our primary source of dynamical (velocity) data in the horizontal? The basic problem is that even aircraft on scientific missions cannot maintain perfectly "flat" trajectories. In a 2D turbulence, the vertical would be smooth and there would be no biases in estimates of spectral exponents. Similarly, in 3D isotropic turbulence, even if the trajectories are non-smooth there is a single exponent (independent of direction), so again the measured exponent can plausibly be taken at face value. However, if the turbulence is anisotropic with different vertical and horizontal exponents – then the aircraft can have fractal trajectories implying long-range correlations between the aircraft position and atmospheric structures, leading to possible biases in the exponents. In addition, even very small mean vertical slopes can lead to statistics being dominated by vertical rather than horizontal exponents. Both of these effects have been found in stratospheric flights which can have both 1.56 dimensional trajectories [*Lovejoy, et al.*, 2004], and then at scales > 300km, Bolgiano-Obukhov ($k^{-11/5}$) rather than Kolmogorov spectra ($k^{-5/3}$; k is a wavenumber). This has made it possible for the first time to explain the major aircraft campaigns (GASP [*Nastrom and Gage*, 1983], MOZAIC [*Lindborg and Cho*, 2001]). This sets the ground for a clear understanding of the horizontal structure.

3 Wide Range Scaling

Scale invariance – no matter how theoretically appealing – would not be of general geophysical interest were it not for the basic empirical fact that geofields display wide range scaling. Figs. 18.1, 18.2 show spectral analyses of visible cloud radiances covering the range of roughly 1 m to 5000 km, and Fig. 18.2 of topography covering the range 1 m to 20,000 km. In both cases, the spectra are "isotropic" i.e. the squared

modulus of the 2D fourier transform has been angle integrated (and in the case of the cloud data, averaged over the available data). In solid earth geophysics it is more usual to angle average the spectrum; in 2D this will reduce the exponent by one. Below we use angle integration since for isotropic processes, the resulting exponents are independent of the dimension of the analysis space. The use of energy spectra has the advantage of relying on familiar data analysis techniques which are quite sensitive to breaks in the scaling. The angle integration "washes out" much of the anisotropy and explains why – in spite of highly diverse cloud morphologies – the isotropic spectra ($E(k)$) can approximately be of the power law form:

$$E(k) \approx k^{-\beta} \tag{1}$$

(k is the modulus of the wavevector $\underline{k}$). The exponents of such power law spectra are scale invariant because they are invariant under the scale change $\underline{k} \to \lambda \underline{k}$ (corresponding in real space to the scale reduction $\underline{x} \to \lambda^{-1} \underline{x}$ where $\underline{r}$ is position vector); the spectra – which keeps its form but which changes by the factor $\lambda^{-\beta}$ - is called "scaling". In physics the term "scaling" is generally reserved for invariance under space-time scale transformations, although occasionally it is also used to describe the tails of algebraic probability distributions, (in this case it refers to scaling in a probability space; see the discussion of SOC below). In the geosciences there is an unfortunate tendency to use "scaling" to denote the general problem of changing from one scale to another even if there are no conserved properties; below we reserve the term for the more precise physics sense referring to invariant properties under (possibly anisotropic) scale changes.

Geodata are frequently scaling in time as well as space: i.e. in *space-time*. Indeed, geofluids generally have well-defined space-time relations so that structures of a given spatial extent live for a statistically well-defined duration. For example in turbulence – and in the atmosphere – the lifetime of a structure ("eddy") is referred to as the "eddy-turn over time". The lifetime corresponding to structures of planetary extent (the "synoptic maximum", [*Kolesnikov and Monin*, 1965]) is about 2 weeks and is thus the natural time scale separating the weather from the climate; the latter being the result of evolution over many eddy turn over times, the former over a single turn-over time. We may therefore expect the long time behaviour of the atmosphere to have different scaling properties from the short time behaviour. The scaling in space and time will generally be anisotropic (section 9), but will also respect the causality principle [*Marsan et al*., 1996], it allows us to exploit the scaling to study predictability issues and to make stochastic forecasts [*Schertzer and Lovejoy*, 2004].

In Fig. 18.3a, we demonstrate the climate scaling using a mean temperature surrogate (from the Greenland GRIP core), the O^{18}/O^{16} ratio showing a clear turbulent-like signal going back over 100,000 years (including highly intermittent "Dansgaard events", the sharp high frequency spikes).

Figure 18.3b shows that the spectrum is also scaling; with exponent only a little smaller than for turbulent temperatures ($\beta \approx 1.4$ compared to 5/3). In fact, it seems that the scaling goes back to about 40,000 years [*Lovejoy and Schertzer*, 1986] after

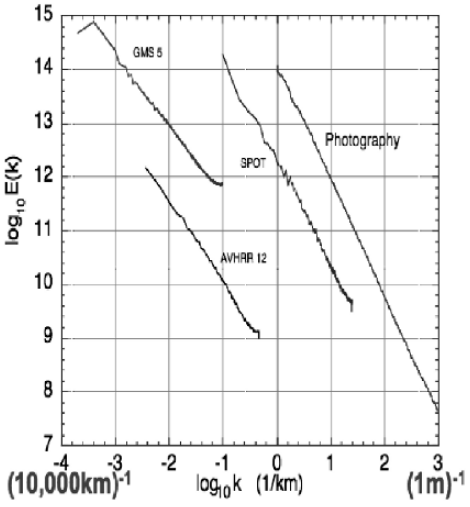


Fig. 18.1. The average spectra (displaced in the vertical for clarity) obtained for GMS, NOAA 12, data as well as a SPOT image over France. The GMS spectrum shows a range of scales of 5120–10 km, the NOAA 12 spectrum shows a range from 256 to 2 km and the SPOT spectrum shows a range from 10 km to 40 m. At the far right we also show the average of 38 ground based pictures (some with resolutions of 50 cm, this is the average of the data discussed in [*Sachs, et al.*, 2002]). Reproduced from [*Lovejoy and Schertzer*, 2006].

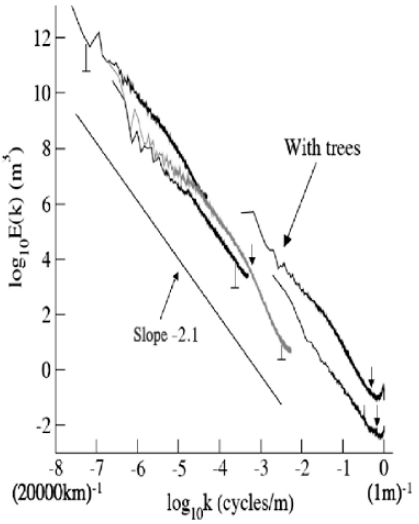


Fig. 18.2. Log/log plot of the spectral energy for four Digital Elevation Maps (DEMs). From right to left: Lower Saxony (with trees, top), Lower Saxony (without trees, bottom), the U.S. at 90 m (in grey), at 30" (about 1km, GTOPO30) and the earth (including bathymetry) at 5' (about 10km), ETOPO5. A reference line of slope −2.10 is on the graph to show the overall slope of the spectra. The small arrows show the frequency at which the spectra are not well estimated due to the inadequate dynamical range of the data; see [*Gagnon, et al.*, 2006] for this theoretical estimate (for ETOPO5, it is well estimated over the whole range). The "semi error bar" symbols indicate the amount of offset due to the resolution dependent factor $\lambda^{K(2)}$ (see [*Gagnon, et al.*, 2006] for this necessary resolution dependent correction). Reproduced from [*Gagnon, et al.*, 2006].

which the spectrum flattens out. Note that there is no strong (above background) signal from the precession of the earth's axis so this analysis tends to make the Milankovitch theory for the ice ages less convincing. Since temperature is the basic climate variable, this suggests that scaling is an appropriate framework for modeling and understanding climate.

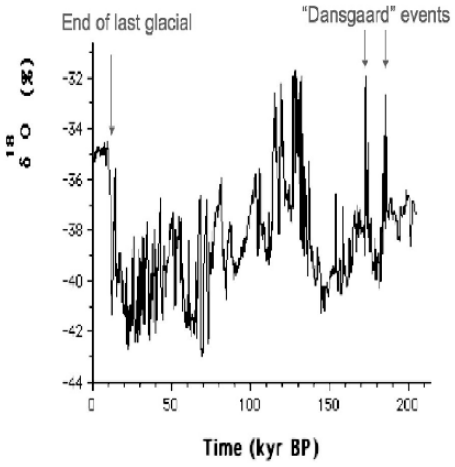


Fig. 18.3a. This figure shows O^{18}/O^{16} in parts per thousand for the Greenland GRIP ice core at 200 year resolution; this is a standard temperature surrogate. The present is at the origin, the past is to the right. Note that the above corresponds to the initial dating of the core; the more modern dating (nonlinearly) shortens the time scale by nearly a factor of 2 (however, it does not much affect the scaling exponents, F. Schmitt private communication). The "Dansgaard events" are the equivalent of going in and out of an ice age in perhaps as little as 100 years. Adapted from [*Schmitt, et al.*, 1995].

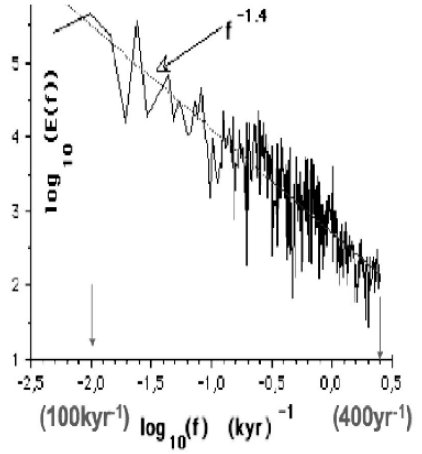


Fig. 18.3b. The spectrum $E(f)$ of the temperature surrogate (Fig 18.3a), with reference slope $f^{1.4}$, this is only a little less steep than the $f^{-5/3}$ spectrum observed in the atmosphere at weather scales (f is the frequency). Adapted from [*Schmitt, et al.*, 1995].

4 From Monofractal Sets to Multifractal Fields

Using fourier spectra, we have seen that the basic atmospheric and earth surface fields display wide range scaling in space and time. Spectra were first widely used to characterize turbulence, and in the early 1970's in conjunction with the development of quasi-gaussian statistical closure models, the theoretical or empirical determination of the spectral exponent became a key task. In quasi-gaussian processes (essentially the "fractional Brownian motions" generalizations of Brownian motion), there is a single basic exponent hence the spectral exponent is simply related to the (unique) fractal dimension of exceedence sets (the set of points exceeding a fixed threshold). However, by the early 1980's, the development of cascade models to study turbulent intermittency lead to the realization that in general the different moments have different scaling behaviors, they are "multiscaling". In fact, the generic result of a cascade process is that the cascade quantity at resolution ε_λ has the statistics:

$$\left\langle \varepsilon_\lambda^q \right\rangle = \lambda^{K(q)} \tag{2}$$

where $K(q)$ is *nonlinear* and convex the moment scaling function and $\lambda = L_{outer}/L$ is the ratio of the largest (outer) cascade scale (L_{outer}) and the scale of observation (L). The symbol "ε" is used for the turbulent (scale by scale) energy flux.

Viewed from the point of multiscaling, spectra are second order statistics (for statistically stationary processes, they are the fourier transforms of the autocorrelation functions; the Wiener-Khintchin theorem), so that the spectrum is only a very partial description of its properties. A rather intuitive and still geometric, but not yet consistent understanding of multifractals can be achieved with the help of thresholds (T) to convert fields $\varepsilon(\underline{x})$ into exceedence sets ($\underline{x}$ is a position vector), and then the use of box-counting to systematically degrade the resolution of the sets, determining the fractal dimension using the formula:

$$N_T(L) \propto L^{-D(T)}; \quad P_T(L) \approx N_T(L)/L^{-d} \approx L^{c(T)}; \quad c(T) = d - D(T) \tag{3}$$

where $N_T(L)$ is the number of LXL sized boxes needed to cover the set of points satisfying $\varepsilon(\underline{x})>T$. Since L^{-d} is the total number of boxes in the space at resolution L, P_T is the probability that a box (size L) placed at random on the set will cover part of the set. $D(T)$ is the dimension and $c(T)$ is the codimension function which is thus a probability exponent. Since probability exponents can be defined without reference to the embedding space of the process (i.e. whether it occurs in a 1-D, 2-D… or for stochastic processes, in infinite D probability spaces), codimensions are generally needed for stochastic fractals and multifractals.

When this "functional" box-counting was done for the topography (Fig. 18.4) or radar reflectivities of rain (Fig. 18.5), it was found that the scaling was excellent: the power law eq. 3 was accurately obeyed for all T, L. However $D(T)$ systematically decreases with threshold, it is not constant as assumed in the monofractal models. Indeed, from the point of view of multifractals, it would have been a miracle if for each threshold T, each (different) set had exactly the same fractal dimension.

It is worth mentioning that the functional box-counting results have important consequences for classical geostatistics (e.g. [*Matheron*, 1970]) which assume (explicitly or implicitly) that geomeasures such as the areas of the topography exceeding a threshold are regular with respect to Lebesgue measures. If this assumption were true, then the areas above a given threshold T would be well-defined independently of the resolution L, i.e. the expression $L^2N_T(L)$ would be independent of L; however since $D(T)<2$ we see that generally it vanishes as $L->0$. Ultimately at millimetric or smaller scales, the scaling will break down yielding a finite limit of $L^2N_T(L)$. However this value will depend on the very small scale details; at any larger resolutions the result will be subjective depending on the observing resolution L.

To go beyond geometry and obtain a consistent formalism, one has to consider scaling fields and thresholds [*Schertzer and Lovejoy*, 1987], i.e. $T \propto \lambda^\gamma$ where the "singularity" γ is arbitrary then we obtain:

$$\Pr(\varepsilon_\lambda > T) = \Pr(\varepsilon_\lambda > \lambda^\gamma) \propto \lambda^{-c(\gamma)} \tag{4}$$

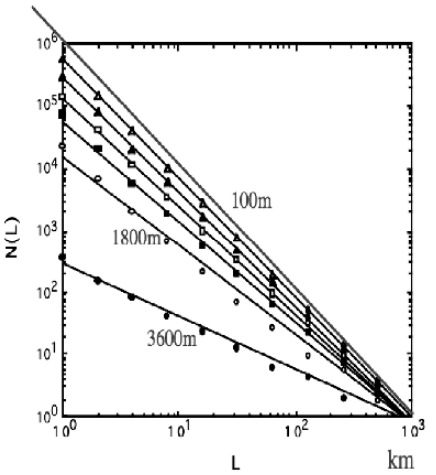


Fig. 18.4. Functional box-counting on French topography data at 1km resolution. For each threshold, the scaling is quite accurate, but as the threshold increases, the slope systematically decreases so that the topography is apparently not monofractal. The line with slope -2 is shown since this is the theoretical assumption of classical geostatistics. Adapted from [*Lovejoy and Schertzer*, 1990].

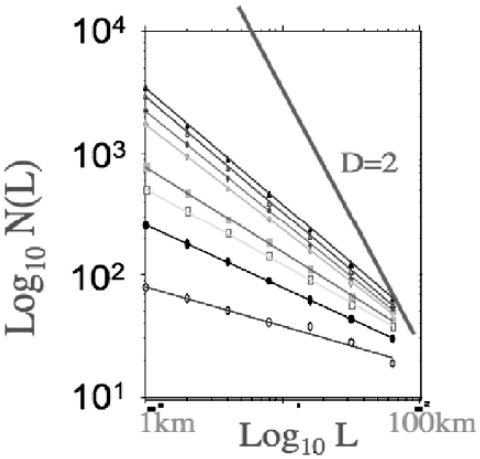


Fig. 18.5. Functional box-counting on radar reflectivity data of rain; the data taken from a weather radar in Montreal, Canada. Each line corresponds to a reflectivity factor increasing by a factor of about 2.5 (starting at the top which is the lowest detectable signal). Although all the different levels are accurately power laws (scaling), the more and more intense rain regions (lower curves) have lower and lower slopes, again we conclude that rain is multifractal. Again, the geostatics theory slope 2 is shown for reference. Adapted from [*Lovejoy, et al.*, 1987].

where "Pr" indicates "probability". Since the moments (eq. 2) are integrals over the probability density (*dPr*), $c(\gamma)$ determines $K(q)$; indeed, the relationship can generally be inverted; with help of the (inverse) Mellin transform (Schertzer, Lovejoy et al. 2002). For large λ, the Mellin transform reduces to the (involutive) Legendre transform:

$$\left\{c(\gamma) = \max_q \left(q\gamma - K(q)\right); \quad K(q) = \max_\gamma \left(q\gamma - c(\gamma)\right)\right\} \tag{5}$$

showing that there is a one-to-one relation between the orders of moments q, and orders of singularity γ and between $c(\gamma)$, $K(q)$ [*Parisi and Frisch*, 1985].

5 Cascades and Data Analyses

The nonlinear terms in the Navier-Stokes equations ruling 3D turbulence conserve the scale by scale (fourier space) flux of energy from large to small scales. In addition, structures interact most strongly with other structures of nearly similar size. Finally, over a wide range, there is no characteristic scale in the nonlinear mechanism. This fourier conservation, fourier "locality" and scale invariance imply a cascade phenomenology, the basis for phenomenological cascades models. Starting with [*Novikov and Stewart*, 1964], [*Yaglom*, 1966, *Mandelbrot*, 1974] these models were developped to study the effects of intermittency. They start with an initially uniform large scale which is successively subdivided into smaller and smaller sub-structures, each multiplicatively modulating the energy flux from the larger scales, with the process repeating scale after scale until in turbulence, viscosity eventually cuts off the cascade (for the atmosphere, at millimeters or less). The resulting cascades have fluxes which respect eq. 2.

In order to empirically test this cascade prediction we may attempt to "invert" the cascade by successively degrading the resolution of the cascade quantities. The only complication is that observable quantities such as the turbulent velocity or the topographic altitude are generally *not* conserved scale by scale. For example in Kolmogorov turbulence, the velocity gradient Δv is related to the conserved energy flux ($\varepsilon_{\Delta x}$) at resolution Δx by:

$$\Delta v = \varepsilon_{\Delta x}^{a} \Delta x^{H}; \quad a = 1/3; \quad H = 1/3 \tag{6}$$

The usual interpretation of this formula (which is essentially the result of dimensional analysis), is that it expresses the equality of the scaling of the left and right hand sides of the equations, hence, averaging the q^{th} power of eq. 6.:

$$\left\langle |\Delta v|^{q} \right\rangle = \left\langle \varepsilon_{\Delta x}^{qa} \right\rangle \Delta x^{qH} = \Delta x^{\xi(q)}; \quad \xi(q) = qH - K(qa) \tag{7}$$

where "<.>" means statistical averaging, and $\xi(q)$ is the (generalized) "q^{th} order structure function exponent". We see that to determine the $K(q)$ exponent characterizing the flux, we must either remove the Δx^{H} term in eq. 6 to estimate ε^{a}, or we may use the structure function to directly determine $\xi(q)$ from eq. 7 and then remove the linear contribution qH. Since power law filtering of ε^{a} by k^{-H} in fourier space has the effect of introducing the linear scaling term Δx^{H} (see section 8), this suggests that to obtain ε^{a} from Δv, that we invert the integration by fractionally differentiating by an order H. It turns out that it suffices to fractionally differentiate with order > H so that for H<1 (the usual case), a standard finite difference approximation (such as in 1-D taking the absolute differences at the finest resolution, or in 2-D taking the modulus of the finite difference gradient vector), is adequate [*Lavallée, et al.*, 1993]. Figs. 18.6, 18.7, show the results on global atmospheric and topographic data. We see clearly that eq. 2 is accurately obeyed over the entire available ranges of scale, and that the effective outer scale (i.e. the scale where the cascade must have started if it is the only source of variability), is of the order of planetary scale (and in Fig. 18.6 slightly larger indicating that there are other sources of

variability at this scale). The figures show that the variability of both the weak (low q) and strong (high q) fluctuations at all the observed scales can be accurately accounted for by multiplicative cascade processes.

6 Multifractal Universality

We have already mentioned that the generic result of cascade processes is that the conserved flux obeys eq. 2; i.e. it involves the convex function $K(q)$. At this level the theory effectively involves an infinite number of parameters ($K(q)$). If no simplifications were to occur, this would be unmanageable. However, as is usually the case in physics when processes are sufficiently iterated or when interactions are sufficiently numerous, we expect that only some of the details will be important.

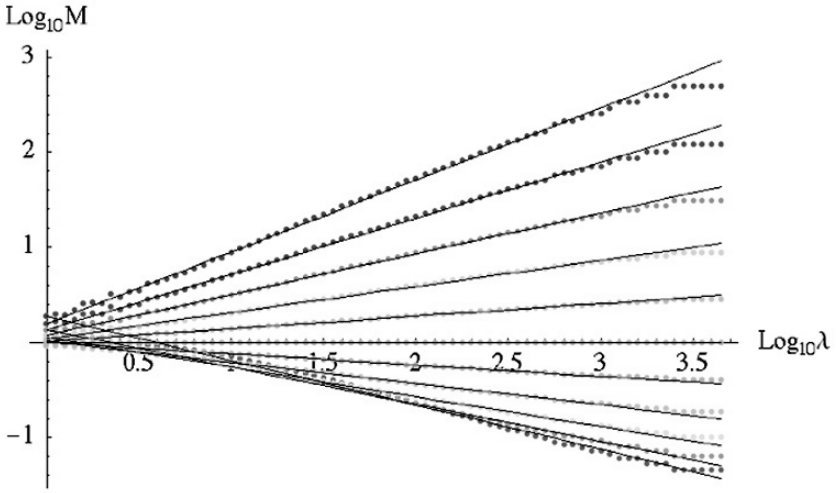


Fig. 18.6. The statistical moments (M) of order q of the gradients of satellite radar reflectivities (Z) from the Tropical Rainfall Monitoring Mission (TRMM) covering the full planetary scale down to 4.3 km, normalized so that the mean = 1. With the exception of very low order moment dominated by (nominally) zero rain reflectivity, the multiscaling holds remarkably well. The data is from 1166 consecutive orbits (70 days) at 4.3km resolution, swath width 220km; the curves from bottom to top are for moments of order q =0, 0.2, 0.4, 0.6… 2.0. Analysis shows that $K(q)$ satisfies the universal form (eq. 8) with α=0.5, C_1=0.63. Multifractal modelling shows that the slight curvature in the above for low q is well explained by this simple threshold detection model: the moments of order <2 are determined within 7% over the entire range 20,000 - 4.3km by the two parameters α, C_1, a threshold equal to one half the mean reflectivity, and an "effective external scale" ≈ 30,000km.

This is the general problem of "universality"; for cascades the issue is complicated because of the highly singular small scale limit (see section 7). It is therefore important to consider the issue of universality over a fixed, (finite) range of scales, and only then take the small scale limit (see the debate [*Schertzer and Lovejoy*, 1997], and [*She and Levesque*, 1994] for a weaker "Log-Poisson" universality). The result is a kind of "multiplicative central limit theorem" leading to:

$$K(q) = \frac{C_1}{\alpha - 1}\left(q^{\alpha} - q\right) \tag{8}$$

where $0<\alpha\leq 2$ is the multifractal index, (it is the Levy exponent of the generator) and $0<C_1<d$ is the codimension of the mean [*Schertzer and Lovejoy*, 1987]. When $\alpha = 1$ we have $K(q) = C_1 q \log q_1$ when $\alpha<2$ the above is valid only for $q\geq 0$. This can be roughly understood by taking the logs of the cascade which is the sum of a large number of logarithmic contributions and therefore (if appropriately centred, normalized) is subject to the additive central limit theorem (which is in fact one of the original applications of "universality" well before the idea became a generally recognized physics principle).

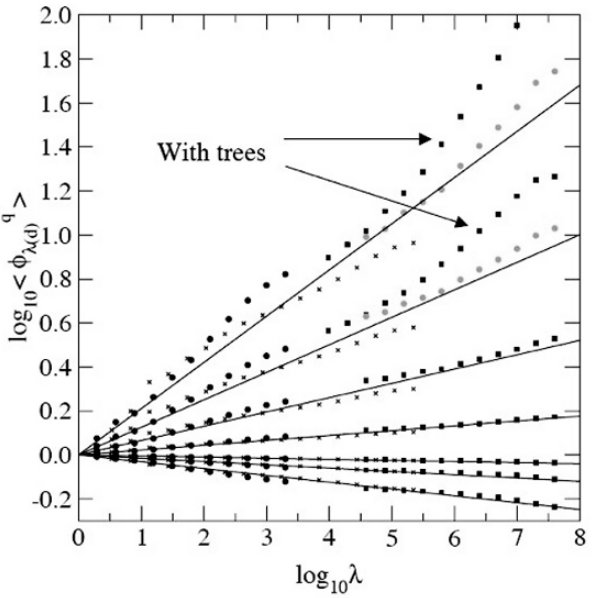


Fig. 18.7. Log/log plot of the normalized trace moments versus the scale ratio λ = *Louter*/*l* (with *Louter* = 20 000 km) for the three DEMs used in Fig. 18.2 (circles correspond to ETOPO5, X's to U.S. and squares to Lower Saxony). The solid lines are there to distinguish between each value of q (from top to bottom, q=2.18, 1.77, 1.44, 1.17, 0.04, 0.12 and 0.51). The moments of the Lower Saxony DEM with trees for q=1.77 and q=2.18 are on the graph (indicated by arrows). The theoretical lines are computed with the global $K(q)$ function i.e. with α=1.79, C_1=0.12, for the parameter estimation and figure, reproduced from [*Gagnon, et al.*, 2006].

The special case α=2 corresponds to the (misnamed) "log-normal" cascades [*Kolmogorov*, 1962], the special case α=0 to the monofractal "β model" [*Frisch, et al.*, 1978]. Many atmospheric and terrestrial surface fields have now been shown to be compatible with eq. 8, and the universality parameters H, C_1, α have been estimated for several dozen geofields (for a review, see [*Lovejoy and Schertzer*, 1995]).

7 Multifractals, Extremes and SOC; the Multifractal Butterfly Effect and Multifractal Phase Transitions

We have mentioned the huge impact of [*Bak, et al.*, 1987]'s paper linking apparently simple avalanche like "sandpile" models to extreme events with algebraic probability distributions. This "classical SOC" has spawned a large number of variants which exploit a cellular-automaton framework with a simple threshold rule applied at a small scale (grid) which leads to both fractal structures and algebraically distributed extremes. What is still under appreciated is the fact that multifractal cascade processes also generate fractal structures with algebraic probabilities: "nonclassical SOC". This effect - as in the usual (chaotic) butterfly effect - is the result of small scale disturbances dominating the large scale and has therefore been called the "multifractal butterfly effect" [*Lovejoy and Schertzer*, 1998b]. It's origin lies in the highly singular small scale cascade limit. This singular behaviour is particularly wild in the general case where the cascade only respects "canonical" (scale by scale) conservation in which it is only the ensemble average energy flux which is conserved scale by scale [*Mandelbrot*, 1974]. In order to avoid the attendant technical difficulties, the simpler but much more restrictive "microcanonical conservation" is all too often considered but it does not have these strong extremes (e.g. the "p model", [*Meneveau and Sreenivasan*, 1987]). In the general canonical cascades, the small scale limit is only well-defined for integrals over finite sets; this leads to the distinction between "bare" cascades quantities (which are the result of the cascade developed over a given range of scales), and the "dressed" properties which are the result of the cascade developed down to infinitely small scales and then averaged up to the same scale. Below a critical order of moment q_D (which depends on the dimension of the averaging space, D), both the bare and dressed properties respect eq. 2; however, for $q>q_D$, the dressed moments (and hence $K(q)$) diverge whereas the "bare" moments converge (for all $q\geq 0$). This implies:

$$\left\{\left\langle \varepsilon_\lambda^q \right\rangle \to \infty; \quad q > q_D \right\} \Leftrightarrow \left\{\Pr\left(\varepsilon_\lambda > s\right) \approx s^{-q_D}; \quad s >> 1 \right\} \tag{9}$$

so that the dressed fluxes display the key feature of SOC: algebraic probabilities. This "nonclassical SOC" [*Schertzer and Lovejoy*, 1994] may often be physically more realistic than classical SOC since whereas the latter is only valid in the "zero-flux limit" (e.g. each sand grain must be added one by one only after the avalanches provoked by the preceding have ended), the multifractal SOC is a generic SOC mechanism valid in systems with quasi-constant fluxes. Note that moments are only infinite in the limit of infinite sample sizes. In fact, both infinite and finite sample size behaviors can be recast in the framework of first and second order multifractal phase

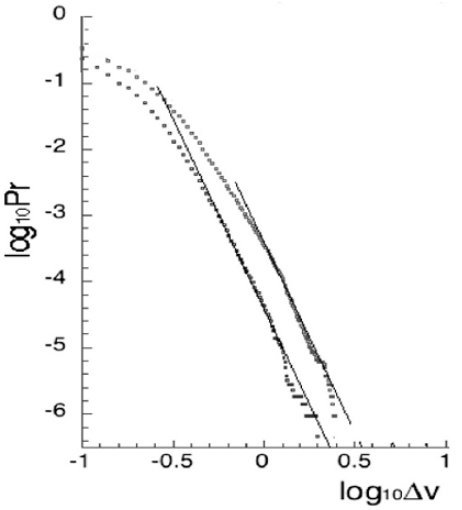


Fig. 18.8. This shows the probability distribution (cumulated from the largest to smallest; Pr ($\Delta v^1 > \Delta v$) for horizontal velocity differences (Δv^1) from the stratospheric ER2 aircraft. The horizontal difference are given for distances of 40m, 80m (left, right), the reference slopes corresponds to eq. 9 with q_D=5.7. This value is in between the value 5 found in the vertical [*Schertzer and Lovejoy*, 1985a] and the temporal value q_D=7 (a surrogate for the horizontal, [*Schmitt, et al.*, 1994]), the in between value may be a consequence of the vertical displacements of the aircraft. See also [*Tchiguirinskaia, et al.*, 2006], [*Tuck, et al.*, 2004]. The data are from 18 aircraft flights, each over paths 1000 -2000 km in length.

transitions [*Schertzer and Lovejoy*, 1992, *Schertzer et al*., 1993], which clarify the different notions of criticality. Fig. 18.8 shows that atmospheric turbulence is indeed apparently a SOC phenomenon (see also [*Chigirinskaya, et al*., 1994; *Schmitt, et al*., 1994] for other atmospheric applications). The extreme nature of multifractals is due to the fact that they can be viewed as random hierarchies of singularities (eq. 4); the structure of the extreme singularities can thus readily stand out. To study this it is helpful to make numerical simulations. Fig. 18.9 shows an example with a log-spiral singularity showing how cyclones can be modeled without needing special "cyclone" models, their "order" apparently emerging from chaos; in the next section we discuss simulations further.

8 The Fractionally Integrated Flux Model and Multifractal Simulations

To model scaling processes, it is natural to use combinations of scale invariant basis functions, i.e. mathematical singularities. Let's first consider the basic additive scaling process, fractional Brownian motion (fBm) and it's generalization, fractional Levy motion (fLm). These can be written as convolutions of noises with power laws i.e. fractional integrations:

$$v(\underline{x}) = \int \frac{\phi_\alpha(\underline{x}')d\underline{x}'}{|\underline{x}-\underline{x}'|^{D-H'}}; \quad H' = H + D/\alpha \tag{10}$$

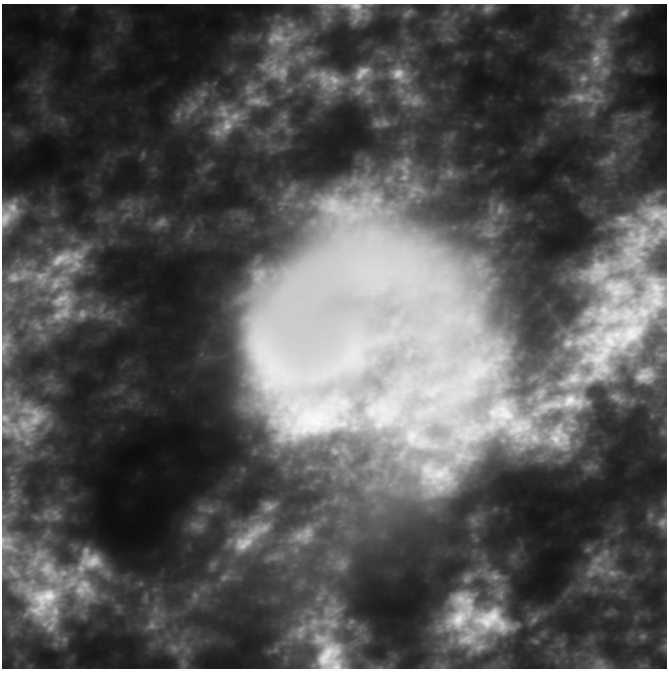


Fig. 18.9. A cyclone emerging from chaos. This multifractal simulation has roughly the observed universal multifractal parameters (α=1.8, C_1=0.1, H=0.33), but is symmetric with respect to a G with complex eigenvalues (hence structures rotate with scale). From time to time, the process produces particularly strong singularities which can dominate the simulation. Here, the event was sufficiently rare that it was "helped" by artificially increasing a single value of the 2^{24} elements of the subgenerator. (A color version of this figure appears between pages 196 and 197).

With $\alpha \leq 2$. ϕ_2 is a gaussian white noise, for $\alpha<2$, ϕ_α is a Levy noise made of uncorrelated Levy random variables and H' is the order of fractional integration. Introducing the fluctuation $\Delta v(\Delta x)=v(x+\Delta x)-v(x)$, the resulting v field has statistics obeying:

$$\left\langle \left|\Delta v\left(\underline{\Delta x}\right)\right|^q \right\rangle \approx \left|\underline{\Delta x}\right|^{\xi(q)}; \quad \xi(q) = qH \tag{11}$$

when $\alpha<2$, $\xi(q)$ diverges for $q>\alpha$. These models are monofractal because the fractal codimension of any level set $v(\underline{x})=T$ has a codimension $c(T)=H$ (i.e. independent of T).

Figure 18.10 shows a comparison of fBm, fLm and a universal multifractal process with the same H value. The multifractal process can also be modeled as a fractional integration of a noise, although this time, the noise is a conservative (cascade) multifractal (a "flux", hence the name Fractionally Integrated Flux, FIF model (see [*Schertzer, Lovejoy et al*., 1997] for its relationships with the renormalization group approach)):

$$v_\lambda(\underline{x}) = \int \frac{\varepsilon_\lambda(\underline{x}')d\underline{x}'}{|\underline{x}-\underline{x}'|^{D-H}}; \quad \varepsilon_\lambda(\underline{x}') = e^{\Gamma_\lambda(\underline{x}')}; \quad \Gamma_\lambda(\underline{x}) \propto C_1^{1/\alpha} \int_{1<|\underline{x}'|<\lambda} \frac{\phi(\underline{x}')d\underline{x}'}{|\underline{x}-\underline{x}'|^{D-H'}} \tag{12}$$

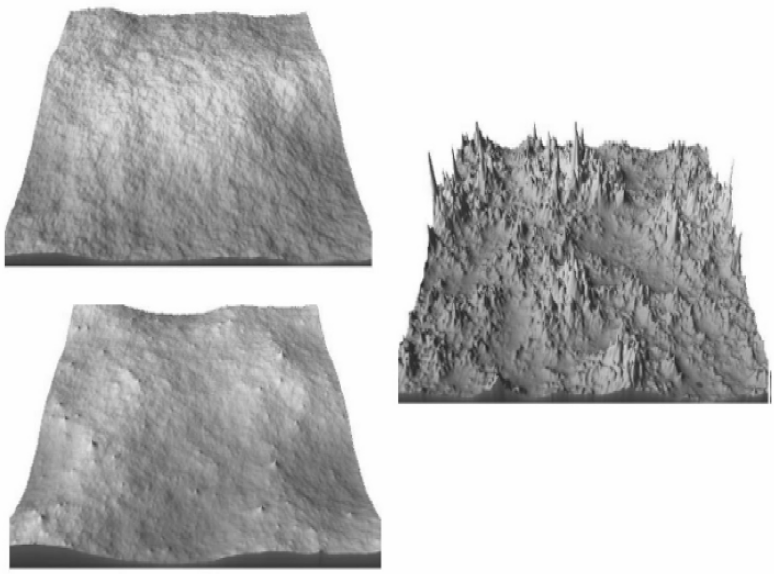


Fig. 18.10. The upper left simulation shows fBm, with H=0.7, lower left fLm with H=0.7, α =1.8, and the right the Multifractal FIF with H=0.7, α =1.8, C_1=0.12 (close to observations for topography, see [*Gagnon, et al.*, 2006]). For more examples of multifractal simulations, see the multifractal explorer site: http://www.physics.mcgill.ca/~gang/multifrac/index.htm. (A color version of this figure appears between pages 196 and 197).

where for $\alpha<2$, ϕ_α is a maximally skewed Levy noise process and $H'=D(1-1/\alpha)$. The resulting ε_λ is multiplicative because it is an exponentiation of the additive generator Γ_λ. From Fig. 18.10 we can see that the fBm gives a relatively uninteresting texture. fBm is fairly limited in its possibilities since due to the central limit theorem (the gaussian special case), a process with the same statistical properties can be produced by using singularities of quite different shapes; it is insensitive to the shape. The fLm on the contrary has extremes which are too strong; as one can see, several strong mountain peaks stand out; in fact, the strong peaks are too strong. Although far from Gaussian, real topography empirically seems to have finite variance (i.e. the probability density tail falls off faster than x^{-3}), so fLm is not a good model. Finally, the multifractal simulation has much more interesting structures, however we are missing the interesting ridges, valleys and other anisotropic features of real geomorphologies.

Considering the "universal multifractals" v_λ, defined by eq. 12 we see that they are isotropic (the singularities have no preferred directions, they depend only on the vector norm), they are therefore "self-similar"; "zoomed" structures will (on average) resemble the unzoomed ones. In addition, they depend on three parameters: the H in eq. 12, and the α, C_1 which define the statistics of the generator Γ_λ. While the parameter H is the order of fractional integration and quantifies the degree of scale invariant smoothing, the qualitative effects of the codimension of the mean (C_1) and the multifractal index (α) are less easy to see. We therefore performed

multifractal simulations (Fig. 18.11) which systematically show the morphologies of the structures obtained by varying these three parameters. The simulations use false colors and each has the same initial random "seed" so that the basic structures are the same. For reference, note that the empirically most common values of α are in the range 1.5-1.8 (the latter being appropriate for topography and cloud radiances, the

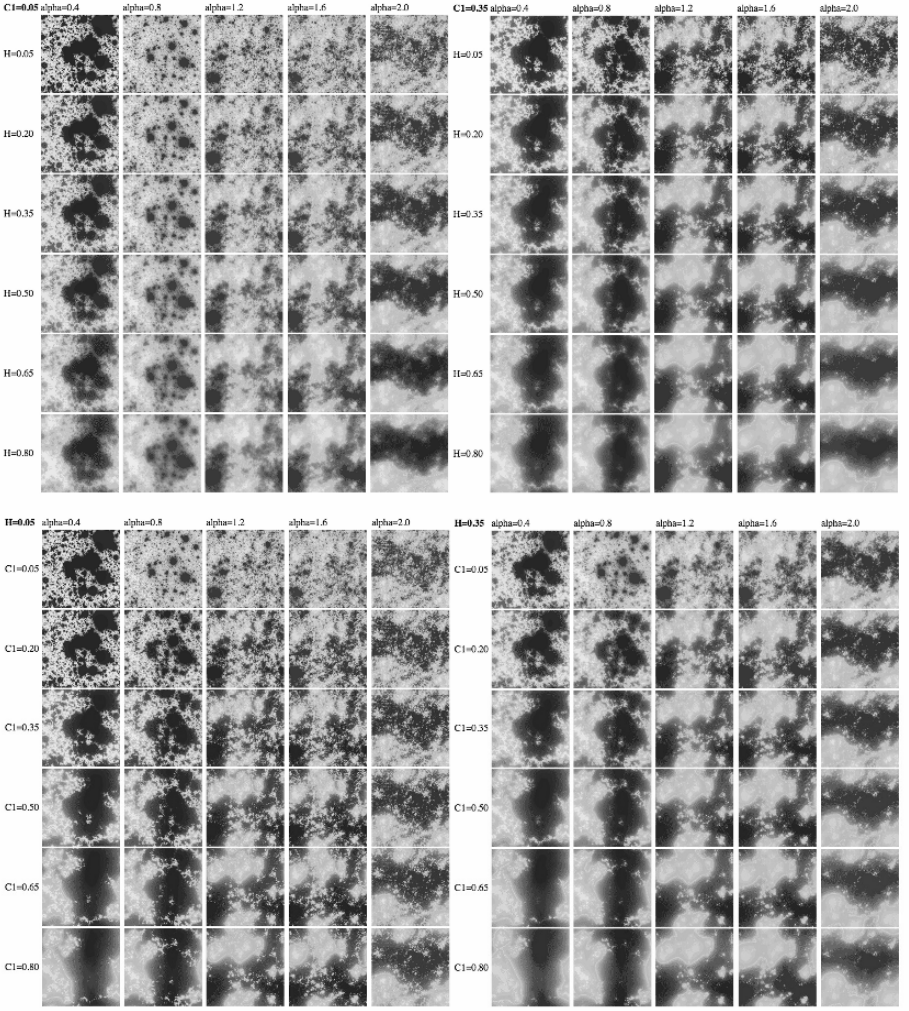


Fig. 18.11. This shows the effect of varying α, C_1, H values on multifractal simulations. The upper figures show the effect of increasing α (left to right, 0.4, 0.8, …2.) and H (top to bottom 0.05, 0.2, …0.8) with C_1 fixed (=0.05 left, 0.35 right). The lower figures show the effect of varying α (left to right, 0.4, 0.8, …2.) and C_1 (top to bottom 0.05, 0.2, …0.8) with H fixed (=0.05 left, 0.35 right). (A color version of this figure appears between pages 196 and 197).

former, for rain and atmospheric turbulence). The parameter C_1 is often fairly low (e.g. in the range 0.05-0.15 for the wind, cloud radiances, topography), although it can be large (0.25-0.7) for rain and turbulent fluxes. While the basic Kolmogorov value of H is 1/3, many fields (such as cloud radiances) are near this, while rain is nearly zero, topography is in the range 0.45-0.7. From Fig. 18.11 we can see that high values of C_1 lead to fields totally dominated by one or two strong structures, while low α values lead to fields dominated by "Levy holes": large regions with extremely low values.

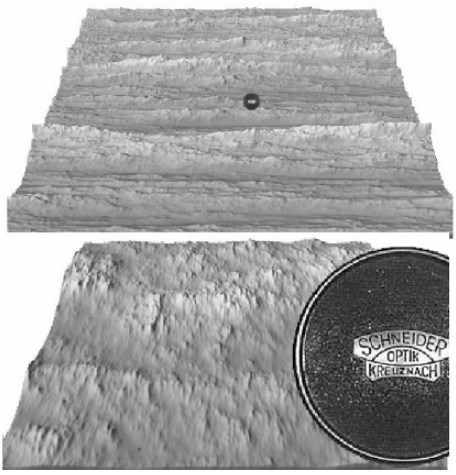


Fig. 18.12a. This self-affine simulation illustrates the "phenomenological fallacy" since both the top and bottom look quite different while having the same generators (G is diagonal with elements 0.8, 1.2), same (anisotropic) statistics at scales differing by a factor of 64 (top and bottom blow-up). The figure shows the proverbial geologists' lens cap at two resolutions differing by a factor of 64. Seen from afar (top), the structures seem to be composed of left to right ridges, however closer inspection (bottom) shows that in fact this is not the case at the smaller scales. (A color version of this figure appears between pages 196 and 197).

Fig. 18.12b. This shows a multifractal simulation looking horizontally through a horizontally stratified cloud layer with α=1.8, C_1=0.1, H=0.333 of a stratified 3D cloud with G diagonal with elements 1, 5/9. This corresponds to a Kolmogorov scaling in the horizontal ($k^{-5/3}$) but a Bolgiano-Obukhov ($k^{11/5}$) scaling in the vertical. Single scatter radiative transfer was used for the rendering. (A color version of this figure appears between pages 196 and 197).

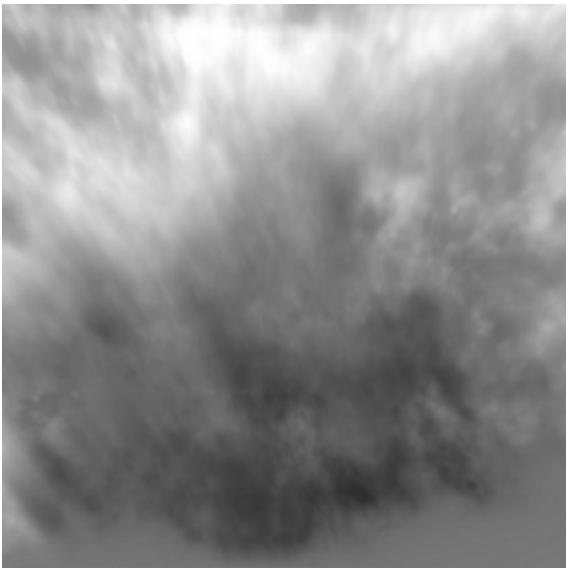


Fig. 18.13. A multifractal simulation of a 3D stratified cloud (with the same G as Fig. 18.12b) with single scattered radiative transfer. Henyey-Greenstein phase functions were used with asymmetry factor g=0.85. (A color version of this figure appears between pages 196 and 197).

9 The Phenenological Fallacy and Generalized Scale Invariance

Geophysicists commonly derive their models from phenomenological classifications based largely on classical (scale bound) notions of scale and shape. Once a phenomenon has been defined - often involving somewhat subjective criterion – models are constructed to explain them. However in this section we shall see that scaling processes – if based on sufficiently strong anisotropic singularities can lead to quite different looking structures at different scales even though the basic mechanism is scale invariant; see Fig. 18.12a for an example and below for a more systematic survey. This possibility demonstrates what we call the "phenomological fallacy" i.e. the danger of inferring process from appearance.

In order to change the shape of the singularities while conserving the basic statistical properties of the process, it turns out to be sufficient to make the replacement everywhere eqs. 10-12:

$$|\underline{\Delta x}| \rightarrow \|\underline{\Delta x}\|; \quad D \rightarrow D_{el} \tag{13}$$

i.e. to replace the usual distance ("$|\ |$") by a "scale function" ("$\|\ \|$") and usual dimension of space by an "elliptical dimension" D_{el} which satisfies the following basic equation scaling:

$$\|T_\lambda \underline{x}\| = \lambda^{-1} \|\underline{x}\|; \quad T_\lambda = \lambda^{-G}; \quad D_{el} = TraceG \tag{14}$$

where T_λ is a scale changing operator which reduces the scale of a vector by a factor λ. In order for the scale function to be scaling (i.e. have no characteristic scale), it must satisfy group properties, hence it must admit a generator *G* as indicated.

Once all the unit vectors $\underline{x}_1$ are specified the nonunit vectors ($\|\underline{x}_\lambda\| = \lambda; \quad \lambda \neq 1$) are then generated by the action of T_λ : $\underline{x}_\lambda = T_\lambda \underline{x}_1$; see [*Schertzer and Lovejoy*, 1985b] for technical details on this Generalized Scale Invariance, GSI. The set of all vectors $\|\underline{x}\| \leq \lambda$ is called a "ball", denoted B_λ; for physical scale functions B_λ must be strictly decreasing (i.e. $B_{\lambda'} \subseteq B_\lambda; \quad \lambda' < \lambda$). We can see that if the replacements $|\underline{x} - \underline{x}'| \to \|\underline{x} - \underline{x}'\|; \quad D \to D_{el}$ are made in the denominators of eqs. 10, 12, with scale functions satisfying eq. 14 then the convolutions will have power law dependencies under "zooming", i.e. the models will be scaling as long as the noises are also scaling (hence the special choices of Gaussian or Levy noise, or in the multifractal case, of multifractal noises).

This generalization of scale, and hence the consequent Generalized Scale Invariance (GSI, [*Schertzer and Lovejoy*, 1985; *Schertzer and Lovejoy*, 1986]) was motivated by the need to deal with the "dimensional transition" that we discussed in the introduction. The main conceptual difficulty was to abandon the classical scheme (which first posits isotropy and then considers scaling). The solution was bold: to first posit scaling, then study the remaining symmetries. To understand the relation between usual distances and generalized scales it is instructive to consider a simple scale function. Consider a (real) 2-D *G* matrix which we diagonalize to yield *G*=diag(*Hx, Hy*) (note that what follows can be generalized to complex eigenvalues or nondiagonalizable matrices). We may now make the following nonlinear transformation:

$$\underline{x}' = (x', y') = \left(\mathrm{sgn}(x)|x|^{1/H_x}, \mathrm{sgn}(y)|y|^{1/H_y}\right) \tag{15}$$

It is then easy to verify that $\left\|\lambda^{-I}\underline{x}'\right\| = \lambda^{-1}\|\underline{x}'\|$ i.e. $\underline{x}$' satisfies the scale eq. 14 but with *G*=*I*=the identity corresponding to the generator of an isotropic scale transformation (sgn (*x*) is the sign of *x*). When *G*=*I* it is easy to verify that a family of solutions of the scale function eq. 14 is:

$$\|\underline{x}'\| = \Theta(\theta')r' \tag{16}$$

where (r',θ') are polar coordinates:

$$r' = |\underline{x}'| = \left(x^{2/H_x} + y^{2/H_y}\right)^{1/2}; \quad \tan\theta' = \frac{y'}{x'} = \frac{\mathrm{sgn}(y)|y|^{1/H_y}}{\mathrm{sgn}(x)|x|^{1/H_x}} \tag{17}$$

and Θ is an arbitrary positive function which specifies the shape of the unit ball (i.e. those vectors with $\|\underline{x}'\| = 1$; their polar equation is r'=1/Θ(θ')) it determines the "trivial" anisotropy (the nonscaling part).

When scale functions are used as the basic singularities, the shapes can be extremely varied, hence demonstrating the possibility of modeling geomorphologies in this way. First consider *G* = *I*: the resulting models will be "self-similar" in the sense that their statistics will vary in power law ways under isotropic "zooming" (blow-ups). If *G* is a diagonal matrix, then the singularities $\|\underline{x}\|^{-\gamma}$ (where γ is the order

of singularity, see eq. 4) are quite different in different directions and the resulting fractals/multifractals are "self-affine" (Fig. 18.12 and 18.14a bottom row). The case where G is nondiagonal and the eigenvalues are real is a generalization in which the main stretching/shrinking occurs along fixed nonorthogonal eigendirections (see Fig. 18.12, 18.14). When the eigenvalues are complex, the eigenvectors rotate continuously as functions of scale, giving rise to spiral type singularities, see Fig. 18.9.

In order to systematically see the effect of varying G, we can parametrize it as follows:

$$G = \begin{pmatrix} d-c & f-e \\ f+e & d+c \end{pmatrix} \tag{18}$$

Defining $a^2 = c^2 + f^2 - e^2$ we see that G has eigenvalues $d+a$, $d-a$, As indicated above, real a is "stratification dominant" (structures rotate no more than once with scale) whereas imaginary a is "rotation dominant" structures rotate an infinite number of times with scale. In order to explore the parameter space, we note that there are 3 rotational invariants: d, a, $(f^2+c^2)^{1/2}$. Hence for example, taking f=0, and varying c loses no generality (other f values are obtained by rotation). Also, we can fix d=1 since it turns out that we can always replace G by G/d as long as we compensate by simultaneously taking the d power of the scale function) i.e we lose no generality with d=1. The two G parameters we explore are thus c, e. The use of a final parameter k allows us to examine the effect of possibly very nonroundish unit scales (see eq. 16). We define it as:

$$\Theta(\theta') = 1 + \frac{1-2^{-k}}{1+2^{-k}}\cos\theta' \tag{19}$$

With this definition, we see that $k = \log_2 (r'_{max}/r'_{min})$ where r'_{max}, r'_{min} are the maximum and minimum radii of the sphero-scale (in the r' space, but this will be close to the ratio in the r space). k=10 thus corresponds to a unit scale which "mixes" conventional scales over a factor of more than 1000 (2^{10}).

The effect of varying the parameters c, e, k is shown in the multifractal simulations shown in Fig. 18.14. All the simulations have α=1.8, C_1=0.1, H=0.33 (the empirical parameters for clouds), and are simulated on 256X256 grids with the same starting seed so that the differences are only due to the anisotropy (the colours go from blue to white indicating values low to high). For isotropic unit scales (k=0, top row Fig. 18.14a) we see the effect of varying c. On the right we display the contours of the corresponding scale functions.

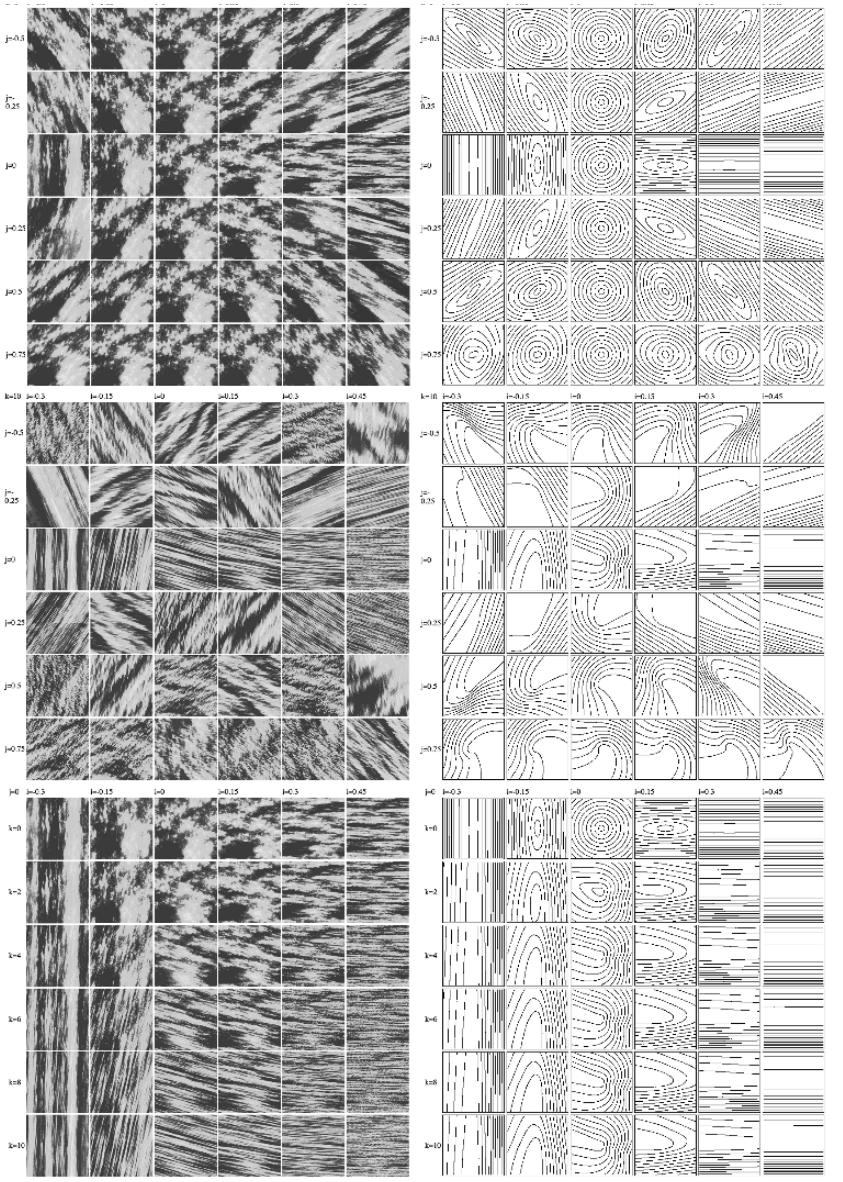


Fig. 18.14a. *Top row* k=0, we vary c (denoted i) from -0.3, -0.15, …0.45 left to right and e (denoted j) from -0.5, -0.25, …0.75 top to bottom. On the right we show the contours of the corresponding scale functions. *Middle row* Same except that k=10. *Bottom row* e=0 the c is increased from -0.3, -0.15, …0.45 left to right, from top to bottom, k is increased from 0,2,4,..10. See text for more details. (A color version of this figure appears between pages 196 and 197).

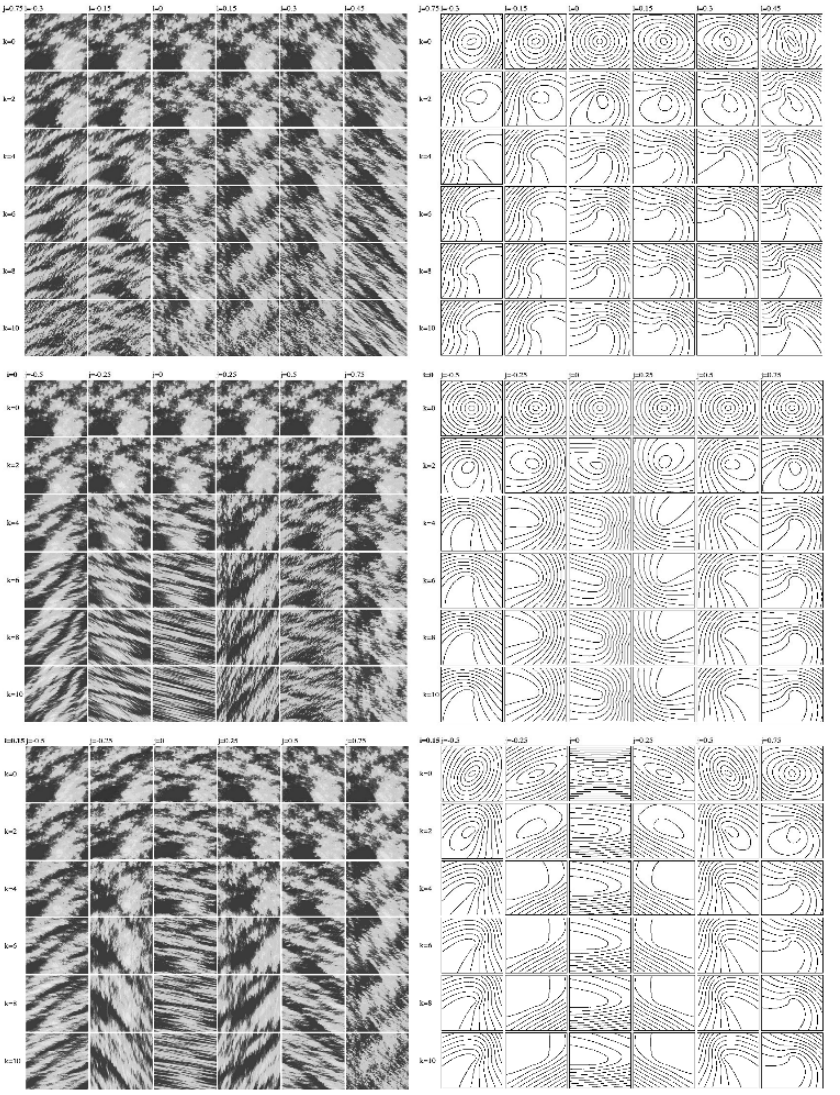


Fig. 18.14b. *Top row* The same as the bottom row of 18.14a except that $e=0.75$. *Middle row* c=0 and e left to right is: -0.5, -0.25, …0.75. *Bottom row* Same as the middle row except that c=0.15. In all rows, from top to bottom, k is increased (0,2,4,..10), the right hand shows the corresponding scale functions. (A color version of this figure appears between pages 196 and 197).

Moving down to the middle row we take k=10 so that the unit ball has a range of scales of 2^{10}; the structures are more filamentary. In the bottom row we take e=0 displaying the effect of varying c: G is thus diagonal, the structures are "self-affine"

(no rotation). Fig. 18.14b (top row) is the same as the bottom row of Fig. 18.14a except that e =0.75 showing the effect of rotation. Since $a^2=c^2-e^2<0$ here (f=0), the eigenvectors of λ^{-G} rotate continuously with scale. In the middle row we fix c=0 and vary e, the bottom row is the same except that c=0.15 so that there is both the effect of stratification (c), and rotation (e). Here the eigenvalues are again complex except in the third column with e=0. Finally, outside our present scope but presumably important for realistic geophysical modeling, we can consider G as a nonlinear operator (rather than a matrix). In this case, the anisotropy depends not only on scale but also on the location. This allows for spatially varying morphologies. In this case, the linear GSI discussed above is simply a locally valid approximation. In this case, the anisotropy depends not only on scale but also on the location. This allows for spatially varying morphologies. In this case, the linear GSI discussed above is simply a locally valid approximation.

10 Conclusions

We are all aware of the extreme variability of geophysical fields over huge ranges of scales: the atmosphere has structures as small as millimeters, and as large as the planet; a ratio of 10^{10}; the surface topography apparently has an even larger ratio. The mathematical modelling of this variability has long stimulated mathematicians and physicists. For example [*Perrin*, 1913] considered the problem of differentiability: "Consider the difficulty in finding the tangent to a point of the coast of Brittany... depending on the resolution of the map the tangent would change. The point is that a map is simply a conventional drawing in which each line has a tangent. On the contrary, an essential feature of the coast is that ... without making them out, at each scale we *guess* the details which prohibit us from drawing a tangent...". The converse problem - integrability ("rectifiability") was considered by [*Steinhaus*, 1954] "... The left bank of the Vistula when measured with increased precision would furnish lengths ten, hundred, and even a thousand times as great as the length read off a school map. A statement nearly adequate to reality would be to call most arcs encountered in nature as not rectifiable. This statement is contrary to the belief that not rectifiable arcs are an invention of mathematicians and that natural arcs are rectifiable: it is the opposite which is true...". [*Richardson*, 1961] quantified integrability by considering the empirical scaling of the coast of Britain and of several frontiers using the "Richardson dividers" method. In his paper [*Mandelbrot*, 1967] "How long is the coast of Britain?" Richardson's scaling exponent was interpreted in terms of a fractal dimension. Later, in his seminal "Fractals: form, chance and dimension" [*Mandelbrot*, 1977] proposed that fractals are ubiquitous in nature. However, when it came to geophysical applications although stimulating, this audacious idea was disappointing: most geofields of interest were mathematical fields (i.e. with a value at (almost) each space-time point such as the atmospheric temperature), and cannot be reduced to geometric sets of points. Furthermore, at that time, the proposed fractal sets were only scale invariant under isotropic scale changes (or occasionally under the only slightly more general self-affine scale changes in

which different exponents act in different orthogonal directions). In a number of fields, by the early 1990's this restrictive isotropic monofractal framework was found to be quite inadequate leading in at least one instance - earth surface morphology – to the virtual abandonment of scaling approaches.

In this paper we have argued that the key to understanding systems with large numbers of degrees of freedom with nonlinear dynamics acting over wide scale ranges was to generalize the notion of scale invariance (and indeed, of scale) in the framework of Generalized Scale Invariance and the basic mathematical object from fractal sets to multifractal functions (or more precisely, densities of measures). While the motivation of the former was to allow the nonlinear dynamics of the system to define the appropriate notion of scale (rather than to impose a priori isotropic, Euclidean notions), the motivation for the latter was to handle nonlinear processes which repeat scale after scale, the generic multifractal process being the cascade.

However, cascades have nontrivial, nonclassical properties, one of them being their highly singular small scale limit which leads to the distinction between "bare" and "dressed" cascade quantities. While the former is the result of the cascade developed from a large scale down to a given (finite) scale, and involves only the larger scale interactions – the latter is the result of integrating ("dressing") a fully developed cascade to the same scale. While the former has long-tailed statistics (e.g. log-Levy, log normal), the latter has fat, "algebraic" tails. If we adopt an operational definition of Self-Organized Criticality as a system with fractal structures and with strong algebraic distributions of extremes, then multifractals generically provide a nonclassical (phase transition) route to SOC.

As seductive as they are, multifractals involve an infinite hierarchy of exponents, and hence would be useless without some robustness principle to eliminate unsteady artifacts. Fortunately, stable, attractive universality classes exist with only three parameters, this enables both compact parametrisations of the exponent functions (and hence the empirical characterization of scaling geofields), it also allows for numerical simulations. In the last part of this paper we show how to make numerical simulations of clouds and topography which are quite realistic both statistically and visually.

During the last twenty years, we have seen the transition from fractal geometry to multifractal processes; we have seen how extremes can be tamed with the help of Self-Organized Criticality, we have witnessed the opening of new areas of applications by generalizing scale invariance from isotropic to anisotropic systems. These developments have in turn paved the way for systematic empirical characterizations of both solid earth and atmospheric geofields, they help overcome longstanding measurement problems (such as the interpretation of aircraft data). They have created the framework needed for realistic modeling of geofields over potentially huge ranges of scale.

11 Acknowledgements

We thank J. S. Gagnon for help with the rendering for Figs. 18.10, 18.11a and E. De Giuili for help with rendering Fig. 18.13 and V. Allaire for data analysis in Fig. 18.6.

12 References

Bak, P., et al. (1987), Self-Organized Criticality: An explanation of 1/f noise, *Physical Review Letter*, *59*, 381-384.

Chigirinskaya, Y., Schertzer, D., Lovejoy, S., Lazarev A., and A. Ordanovich (1994). "Unified multifractal atmospheric dynamics tested in the tropics, part I: horizontal scaling and self organized criticality." *Nonlinear Processes in Geophysics 1(2/3)*: 105-114.

Dewan, E. (1997), Saturated-cascade similtude theory of gravity wave sepctra, *J. Geophys. Res.*, *102*, 29799-29817.

Feigenbaum, M. J. (1978), Quantitative universality for a class of nonlinear transformations, *J. Stat. Phys.*, *19*, 25.

Frisch, U., et al. (1978), A simple dynamical model of intermittency in fully develop turbulence, *Journal of Fluid Mechanics*, *87*, 719-724.

Gage, K. S. (1979). "Evidence for k-5/3 law inertial range in meso-scale two dimensional turbulence." *Journal of Atmospheric Sciences*, *36*, 1979.

Gagnon, J., S., et al. (2006), Multifractal earth topography, *Nonlin. Proc. Geophys.*, *13*, 541-570.

Gardner, C. (1994), Diffusive filtering theory of gravity wave spectra in the atmosphere, *J. Geophys. Res.*, *99*, 20601.

Klinkenberg, B., and M. F. Goodchild (1992), The fractal properties of topography: A comparison of methods, *Earth surface Proc. and Landforms*, *17*, 217-234.

Kolesnikova, V. N., and A. S. Monin (1965), Spectra of meteorological field fluctuations, *Izvestiya, Atmospheric and Oceanic Physics*, *1*, 653-669.

Kolmogorov, A. N. (1962), A refinement of previous hypotheses concerning the local structure of turbulence in viscous incompressible fluid at high Raynolds number, *Journal of Fluid Mechanics*, *83*, 349.

Lavallée, D., et al. (1993), Nonlinear variability and landscape topography: analysis and simulation, in *Fractals in geography*, edited by L. De Cola and N. Lam, pp. 171-205, Prentice-Hall, Englewood, N.J.

Lilly, D., and E. L., Paterson (1983). "Aircraft measurements of atmospheric kinetic energy spectra." *Tellus, 35A*, 379-382.

Lilley, M., et al. (2004), 23/9 dimensional anisotropic scaling of passive admixtures using lidar aerosol data, *Phys. Rev. E*, *70*, 036307-036301-036307.

Lindborg, E., and J. Cho (2001), Horizontal velocity structure functions in the upper troposphere and lower stratosphere i-ii. Observations, theoretical considerations, *J. Geophys. Res.*, *106*, 10223.

Lorenz, E. N. (1963), Deterministic Nonperiodic flow, *J. Atmos. Sci.*, *20*, 130-141.

Lovejoy, S., and D. Schertzer (1986), Scale invariance in climatological temperatures and the spectral plateau, *Annales Geophysicae*, *4B*, 401-410.

Lovejoy, S., and D. Schertzer (1990), Our multifractal atmosphere: A unique laboratory for non-linear dynamics, *Physics in Canada*, *46*, 62.

Lovejoy, S., and D. Schertzer (1995), How bright is the coast of Brittany? in *Fractals in Geoscience and Remote Sensing*, edited by G. Wilkinson, pp. 102-151, Office for Official Publications of the European Communities, Luxembourg.

Lovejoy, S., and D. Schertzer (1998a), Stochastic chaos and multifractal geophysics, in *Chaos, Fractals and models 96*, edited by F. M. Guindani and G. Salvadori, Italian University Press.

Lovejoy, S., and D. Schertzer (1998b), Stochastic chaos, scale invariance, multifractals and our turbulent atmosphere, in *ECO-TEC: Architecture of the In-between*, edited by e. b. A. Marras, p. in press, Storefront Book series, copublished with Princeton Architectural Press.

Lovejoy, S., and D. Schertzer (2006), Multifractals, cloud radiances and rain, *J. Hydrol.*, *322*, 59-88.

Lovejoy, S., et al. (1987), Functional Box-Counting and Multiple Dimensions in rain, *Science*, *235*, 1036-1038.

Lovejoy, S., et al. (2004), Fractal aircraft trajectories and nonclassical turbulent exponents, *Physical Review E*, *70*, 036306-036301-036305.

Lovejoy, S., et al. (2007), Is Kolmogorov turbulence relevant in the atmosphere? *Geophys. Resear. Lett.* (in press).

Mandelbrot, B. B. (1967), How long is the coastline of Britain? Statistical self-similarity and fractional dimension, *Science*, *155*, 636-638.

Mandelbrot, B. B. (1974), Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier, *Journal of Fluid Mechanics*, *62*, 331-350.

Mandelbrot, B. B. (1977), *Fractals, form, chance and dimension*, Freeman, San Francisco.

Marsan, D., Schertzer, D., S. Lovejoy, 1996: Causal Space-Time Multifractal modelling of rain. *J. Geophy. Res. 31D, 26*, 333-26, 346.

Matheron, G. (1970), Random functions and their applications in geology, in *Geostatistics*, edited by D. F. Merriam, pp. 79-87, Plenum Press, New York.

Meneveau, C., and K. R. Sreenivasan (1987), Simple multifractal cascade model for fully develop turbulence, *Physical Review Letter*, *59*, 1424-1427.

Monin, A. S. (1972). Weather forecasting as a problem in physics. Boston Ma, MIT press.

Nastrom, G. D., and K. S. Gage (1983), A first look at wave number spectra from GASP data, *Tellus*, *35*, 383.

Novikov, E. A., and R. Stewart (1964), Intermittency of turbulence and spectrum of fluctuations in energy-disspation, *Izv. Akad. Nauk. SSSR. Ser. Geofiz.*, *3*, 408-412.

Orlanski, I. A rational subdivision of scales for atmospheric processes, *Bull. Amer. Met. Soc.*, *56*, 527-530.

Parisi, G., and U. Frisch (1985), A multifractal model of intermittency, in *Turbulence and predictability in geophysical fluid dynamics and climate dynamics*, edited by M. Ghil, et al., pp. 84-88, North Holland, Amsterdam.

Perrin, J. (1913), *Les Atomes*, NRF-Gallimard, Paris.

Richardson, L. F. (1961), The problem of contiguity: an appendix of statistics of deadly quarrels, *General Systems Yearbook*, *6*, 139-187.

Rodriguez-Iturbe, I., and A. Rinaldo (1997), *Fractal River Basins*, 547 pp., Cambridge University Press, Cambridge.

Ruelle, D. and F. Takens (1971). "On the nature of turbulence." Communications on Mathematical Physics 20 & 23: 167-192 & 343-344.

Sachs, D., et al. (2002), The multifractal scaling of cloud radiances from 1m to 1km, *Fractals*, *10*, 253-265.

Schertzer, D., and S. Lovejoy (1985a), The dimension and intermittency of atmospheric dynamics, in *Turbulent Shear Flow 4*, edited by B. Launder, pp. 7-33, Springer-Verlag.

Schertzer, D., and S. Lovejoy (1985b), Generalised scale invariance in turbulent phenomena, *Physico-Chemical Hydrodynamics Journal*, *6*, 623-635.

Schertzer, D. and S. Lovejoy (1986). Generalised scale invariance and anisotropic inhomogeneous fractals in turbulence. Fractals in Physics. L. Pietronero and E. Tosatti. Amsterdam, North-Holland: 457-460.

Schertzer, D., and S. Lovejoy (1987), Physical modeling and Analysis of Rain and Clouds by Anisotropic Scaling of Multiplicative Processes, *Journal of Geophysical Research*, *92*, 9693-9714.

Schertzer, D. and S. Lovejoy (1992). "Hard and Soft Multifractal processes." *Physica A 185*, 187-194.

Schertzer, D., and S. Lovejoy (1994), Multifractal Generation of Self-Organized Criticality, in *Fractals In the natural and applied sciences*, edited by M. M. Novak, pp. 325-339, Elsevier, North-Holland.

Schertzer, D., and S. Lovejoy (1997), Universal Multifractals do Exist! *J. Appl. Meteor.*, *36*, 1296-1303.

Schertzer, D., S. Lovejoy, 2004: Uncertainty and Predictability in Geophysics: Chaos and Multifractal Insights, In: State of the Planet, Frontiers and Challenges in Geophysics, edited by R. S. J. Sparks, and C. J. Hawkesworth, pp. 317-334, AGU, Washington.

Schertzer, D. and S. Lovejoy (2006). Multifractals en Turbulence et Géophysique. Irruption des Géométries Fractales dans la Sciences. G. Belaubre and F. Begon. Paris, Académie Européeene Interdisciplinaire des Sciences: 189-209.

Schertzer, D., et al. (1993), Generic Multifractal phase transitions and self-organized criticality, in *Cellular Automata: prospects in astronomy and astrophysics*, edited by J. M. Perdang and A. Lejeune, pp. 216-227, World Scientific.

Schertzer, D., S. Lovejoy, F. Schmitt, I. Tchiguirinskaia and D. Marsan (1997). "Multifractal cascade dynamics and turbulent intermittency." *Fractals* 5(3): 427-471.

Schertzer, D., S. Lovejoy and P. Hubert (2002). An Introduction to Stochastic Multifractal Fields. ISFMA Symposium on Environmental Science and Engineering with related Mathematical Problems. A. Ern and W. Liu. Beijing, High Education Press. 4: 106-179.

Schertzer, D., et al. (2002), Which chaos in the rainfall-runoff process? *Hydrol. Sci.*, *47*, 139-148.

Schmitt, F., et al. (1995), Multifractal analysis of the Greenland Ice-core project climate data., *Geophys. Res. Lett*, *22*, 1689-1692.

Schmitt, F., et al. (1994), Empirical study of multifractal phase transitions in atmospheric turbulence, *Nonlinear Processes in Geophysics*, *1*, 95-104.

She, Z., S., and E. Levesque (1994), *Phys. Rev. Lett.*, *72*, 336.

Steinhaus, H. (1954), Length, Shape and Area, *Colloquium Mathematicum*, *III*, 1-13.

Tchiguirinskaia, I., et al. (2006), Wind extremes and scales: multifractal insights and empirical evidence, in *EUROMECH Colloquium on Wind Energy*, edited by P. S. Eds. J. Peinke, Springer-Verlag.

Tuck, A. F., et al. (2004), Scale Invariance in jet streams: ER-2 data around the lower-stratospheric polar night vortex, *Q. J. R. Meteorol. Soc.*, *130*, 2423-2444.

Van der Hoven, I. (1957), Power spectrum of horizontal wind speed in the frequency range from.0007 to 900 cycles per hour, *Journal of Meteorology*, *14*, 160-164.

Wilson, K. G. (1971), "Renormalization Group and Critical Phenomena. I. Renormalization Group and the Kadanoff Scaling Picture." *Phys. Rev. B* 4(9): 3174-3183.

Yaglom, A. M. (1966), The influence on the fluctuation in energy dissipation on the shape of turbulent characteristics in the inertial interval, *Sov. Phys. Dokl.*, *2*, 26-30.

19 Statistics of Return Intervals and Extreme Events in Long-term Correlated Time Series

Armin Bunde[1], Jan F. Eichner[1], Jan W. Kantelhardt[1,2] and Shlomo Havlin[3]

[1] Justus-Liebig-Universität Giessen, Germany, Institut für Theoretische Physik III, armin.bunde@physik.uni-giessen.de, jan.f.eichner@physik.uni-giessen.de
[2] Martin-Luther-Universität Halle-Wittenberg, Germany, Institut für Physik und Zentrum für Computational Nanoscience, jan.kantelhardt@physik.uni-halle.de
[3] Bar-Ilan University, Ramat-Gan, Israel, Minerva Center and Department of Physics, havlin@ophir.ph.biu.ac.il

Abstract. We review our studies of the statistics of return intervals and extreme events (maxima) in long-term power-law correlated data sets characterized by correlation exponents γ between 0 and 1 and different (Gaussian, exponential, power-law, and log-normal) distributions. We found that the long-term memory leads (i) to a stretched exponential distribution of the return intervals (Weibull distribution with an exponent equal to γ), (ii) to clustering of both small and large return intervals, and (iii) to an anomalous behavior of the mean residual time to the next extreme event that increases with the elapsed time in a counterintuitive way. For maxima within time segments of fixed duration R we found that (i) the integrated distribution function converges to a Gumbel distribution for large R similar to uncorrelated signals, (ii) the speed of the convergence depends on both, the long-term correlations and the initial distribution of the values, (iii) the maxima series exhibit long-term correlations similar to those of the original data, and most notably (iv) the maxima distribution as well as the mean maxima significantly depend on the history, in particular on the previous maximum. Most of the effects revealed in artificial data can also be found in real hydro- and climatological data series.

1 Introduction

In recent years there is growing evidence that many natural records exhibit long-term persistence (Bunde and Havlin 1994; Bunde, Kropp, and Schellnhuber 2002). Prominent examples include hydrological data (Hurst, Black, and Simaika 1965; Mandelbrot and Wallis 1969; Rodriguez-Iturbe and Rinaldo 1997; Kantelhardt, Koscielny-Bunde, Rybski, Braun, Bunde, and Havlin 2006; Koscielny-Bunde, Kantelhardt, Braun, Bunde, and Havlin 2006), meteorological and climatological records (Koscielny-Bunde, Bunde, Havlin, and Goldreich 1996; Pelletier and Turcotte 1997; Koscielny-Bunde, Bunde, Havlin, Roman, Goldreich, and Schellnhuber 1998; Tsonis, Roebber, and Elsner 1999; Talkner and Weber 2000; Eichner, Koscielny-Bunde, Bunde, Havlin, and Schellnhuber 2003), turbulence data (Shlesinger, West, and Klafter 1987; Prasad, Meneveau, and Sreenivasan 1988), as well as physiological records (Peng, Buldyrev, Havlin, Simons, Stanley, and Goldberger 1994; Bunde, Havlin, Kantelhardt, Penzel, Peter, and Voigt 2000; Kantelhardt, Penzel, Rostig, Becker, Havlin, and Bunde 2003), and DNA sequences (Peng, Buldyrev, Goldberger, Havlin, Sciortino, Simons, and Stanley 1992; Arneodo, Bacry, Graves, and Muzy 1995).

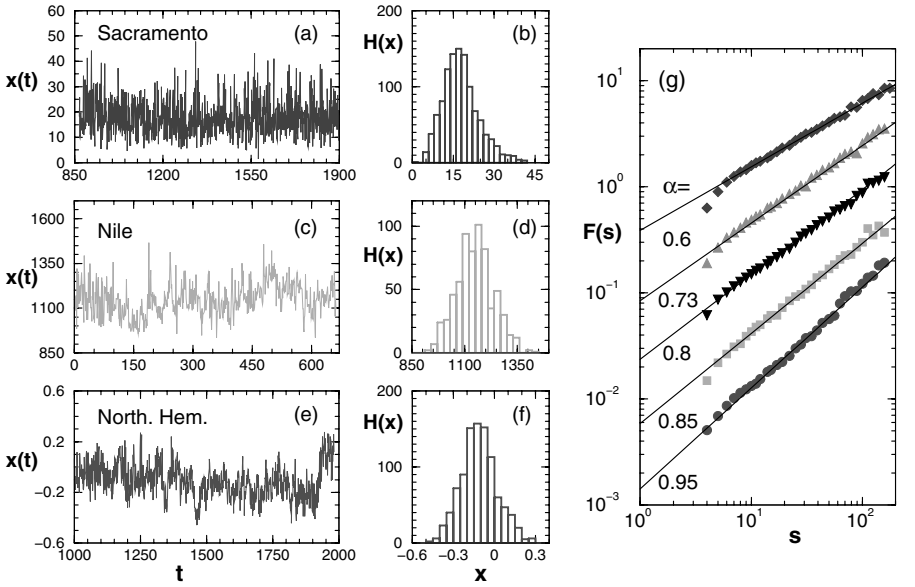


Fig. 19.1. Three annual data series $x(t)$ and their distribution densities $P(x)$; (a,b) reconstructed river run-off of the Sacramento River (USA) (Meko et al. 2001), (c,d) the historical water level minima of the Nile River (EGY) (Whitcher et al. 2002), and (e,f) the reconstructed northern hemisphere temperature (Mann et al. 1999). The data and more information about the data can be achieved from the NOAA Paleoclimatology Program in the internet: http://www.ngdc.noaa.gov/paleo/recons.html. In part (g) the corresponding correlation exponents have been determined. The slopes α of the DFA2 fluctuation function are related to the γ values via $\alpha = 1 - \gamma/2$; the results for the reconstructed temperature record of Baffin Island (Moore et al. 2001) and the reconstructed precipitation in New Mexico (USA) (Grissino-Mayer 1996) are also included. The measured γ values are 0.8 (Sacramento), 0.55 (Baffin), 0.4 (New Mexico), 0.3 (Nile), and 0.1 (Northern hemisphere). From Bunde et al. 2005.

Long-term correlations have also been found in the volatility of economic records (Liu, Cizeau, Meyer, Peng, and Stanley 1997; Yamasaki, Muchnik, Havlin, Bunde, and Stanley 2005) and the internet-traffic (Leland, Taqqu, Willinger, and Wilson 1994; Paxson and Floyd 1995).

In long-term persistent records (x_i), $i = 1, \ldots, N$ with mean $\overline{x}$ and standard deviation σ_x the autocorrelation function $C_x(s)$ decays by a power law,

$$C_x(s) = \frac{1}{\sigma_x^2}\left\langle (x_i - \overline{x})(x_{i+s} - \overline{x})\right\rangle \equiv \frac{1}{\sigma_x^2 (N-s)} \sum_{i=1}^{N-s} (x_i - \overline{x})(x_{i+s} - \overline{x}) \sim s^{-\gamma} \qquad (1)$$

where γ denotes the correlation exponent, $0 < \gamma < 1$. Such correlations are named 'long-term' since the mean correlation time $T = \int_0^\infty C_x(s)\, ds$ diverges for infinitely long series (in the limit $N \to \infty$). Power-law long-term correlations according to Eq. 1 correspond to a power spectrum $P(f) \sim f^{-\beta}$ with $\beta = 1 - \gamma$ according to the Wiener-Khintchin theorem.

Figure 19.1 shows three examples with parts of annual data series exhibiting different degrees of long-term correlations. These records are (a) the reconstructed river run-off of the Sacramento River (USA) (Meko, Therrell, Baisan, and Hughes 2001), (c) the historical water level minima of the Nile River (EGY) (Whitcher, Byers, Guttorp, and Percival 2002), and (e) the reconstructed northern hemisphere temperature (Mann, Bradley, and Hughes 1999). The figure also shows the histograms $H(x)$ of the values x_i of the data; we denote the corresponding distribution density by $P(x)$. To test the records for long-term correlations, we have employed the 2nd order detrended fluctuation analysis (DFA2) (Peng et al. 1994; Bunde et al. 2000; Kantelhardt, Koscielny-Bunde, Rego, Havlin, and Bunde 2001; Eichner et al. 2003). In DFA2, one considers the cumulated sum of the data $Y_j = \sum_{i=1}^{j} x_i$ and studies, in time windows of length s, the mean fluctuation $F(s)$ of Y_j around the best quadratic fit. For long-term correlated data, $F(s)$ scales as $F(s) \sim s^{\alpha}$, with $\alpha = 1 - \gamma/2$. To standardize the records, we subtracted each record by its mean and divided by its variance. The fluctuation functions for the three records from Fig. 19.1(a,c,e) are shown in Fig. 19.1(g) together with those obtained from the reconstructed temperature record of Baffin Island (Moore, Hughen, Miller, and Overpeck 2001) and the reconstructed precipitation in New Mexico (USA) (Grissino-Mayer 1996). The measured γ values are 0.8 (Sacramento), 0.55 (Baffin), 0.4 (New Mexico), 0.3 (Nile), and 0.1 (Northern hemisphere). The observed phenomena should also occur in heartbeat records, internet traffic and stock market volatility and have to be taken into account for an efficient risk evaluation.

In the following we describe how the presence of long-term correlations affects the statistics of the return intervals r between events above a certain threshold value q. We found that the long-term memory in the data leads to a stretched exponential distribution of long return intervals, $\ln P_q(r) \sim -r^{\gamma}$, complemented by a power-law behavior for short return intervals. In short long-term correlated records, however, the stretched exponential can become hard to be observed and to be differentiated from a simple exponential decay. In addition, we found long-term correlations among the return intervals themselves, yielding pronounced clustering of both small and large return intervals. Classical autocorrelation analysis is more suitable for the study of these correlation properties of return interval series than sophisticated DFA approaches. Finally, we observed an anomalous behavior of the mean residual time to the next event that depends on the history and increases with the elapsed time in a counterintuitive way. We present an analytical scaling approach and demonstrate that all these features can be seen in long climate records.

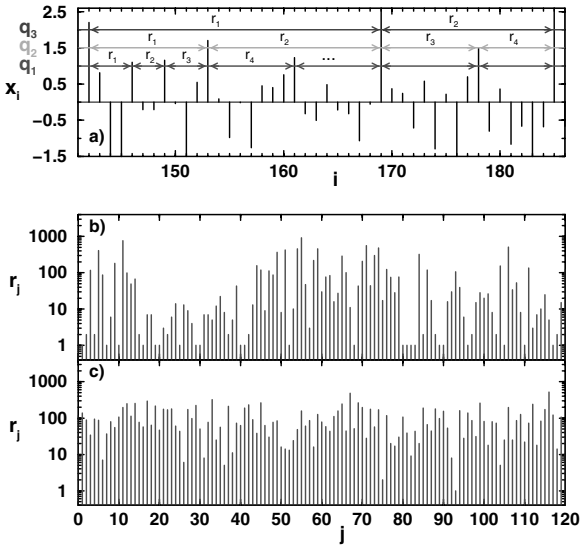


Fig. 19.2. Return intervals: (a) Return intervals r between events x_i above given thresholds q. (b) Sequence of return intervals r of long-term correlated data x_i with $\gamma = 0.4$ for a fixed threshold $q = 2.327$ so that the average r-value $R_q = 100$. Note the logarithmic scale of the r axis. (c) Sequence of return intervals for uncorrelated data (shuffled x_i of (b)). In (b) more epochs with r-values of small and large size appear compared to (c).

The presence of long-term correlations also affects the statistics of the extreme events, i. e., the maxima values of the signal within time segments of fixed duration R. We found numerically that (i) the integrated distribution function of the maxima converges to a Gumbel distribution for large R similar to uncorrelated signals, (ii) the deviations for finite R depend on the initial distribution of the records and on their correlation properties, (iii) the maxima series exhibit long-term correlations similar to those of the original data, and most notably (iv) the maxima distribution as well as the mean maxima significantly depend on the history, in particular on the previous maximum. The last item implies that conditional mean maxima and conditional maxima distributions (with the value of the previous maximum as condition) should be considered for improved extreme event prediction. We provide indications that this dependence of the mean maxima on the previous maximum occurs also in observational long-term correlated records.

2 Statistics of Return Intervals

The statistics of return intervals between extreme events is a powerful tool to characterize the temporal scaling properties of experimental time series and to derive quantities for the estimation of the risk for hazardous events like floods, very high

temperatures, or earthquakes. We focus on long-term correlated signals with different distributions of the values (Gaussian, exponential, power-law, and log-normal distributions) (Eichner, Kantelhardt, Bunde, and Havlin 2007).

2.1 Data generation and mean return interval

In our numerical procedure, we generated sequences of random numbers (x_i) of length $N = 2^{21}$ with either Gaussian, exponential, power-law, or log-normal distribution with unit variance. The corresponding distribution densities $P(x)$ are given by:

$$P_{\text{Gauss}}(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-x^2/2\sigma\right) \tag{2a}$$

$$P_{\text{exp}}(x) = \frac{1}{x_0} \exp\left(-x/x_0\right) \tag{2b}$$

$$P_{\text{power}}(x) = (\delta - 1)\, x^{-\delta} \tag{2c}$$

$$P_{\text{log-norm}}(x) = \frac{1}{\sqrt{2\pi}x} \exp\left[-\left(\ln x + \mu\right)^2/2\right] \tag{2d}$$

We chose, without loss of generality, $\sigma = 1$, $x_0 = 1$, and $\mu = 0.763$ in Eqs. (2a), (2b), (2d) and select $\delta = 5.5$ in Eq. (2c). The long-term correlations were introduced by the Fourier-filtering technique (Turcotte 1992; Makse, Havlin, Schwartz, and Stanley 1996). To preserve the shape of the distribution after retransforming the data from Fourier space we applied an iterative method (Schreiber and Schmitz 1996). For each distribution density $P(x)$ we generated 150 data sets using 1000 iterations, restoring the desired power spectrum by Fourier-filtering and restoring the desired distribution by rank-ordered replacement of the values in each iteration until convergence is achieved. For the Gaussian data, one iteration is sufficient since Fourier-filtering preserves the Gaussian distribution. We tested the quality of the long-term correlations of the data with detrended fluctuation analysis (DFA) (Peng et al. 1994; Bunde et al. 2000) and autocorrelation function analysis.

For describing the recurrence of rare events exceeding a certain threshold q, defined in units of the standard deviations of the original distribution $P(x)$, we investigated the statistics of the return intervals r between these events as illustrated in Fig. 19.2(a) (Bunde, Eichner, Kantelhardt, and Havlin 2003; Bunde, Eichner, Kantelhardt, and Havlin 2004; Bunde, Eichner, Kantelhardt, and Havlin 2005; Eichner et al. 2007;

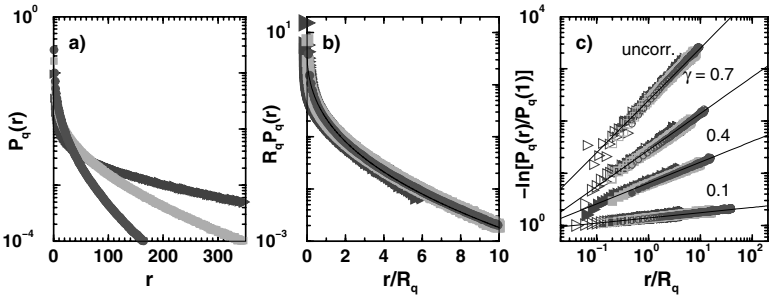


Fig. 19.3. (a) Distribution density $P_q(r)$ of the return intervals r for three thresholds q = 1.5 (R_q ≅ 15, circle), 2.0 (44, square), and 2.5 (161, triangle) for long-term correlated data (γ = 0.4). (b) When plotting $P_q(r)$ multiplied by R_q versus r/R_q, the three curves collapse to a single stretched exponential curve (black line). (c) Double-logarithmic plot of $-\ln[P_q(r)/P_q(1)]$ as a function of r/R_q for γ = 0.1, 0.4, 0.7 as well as for the shuffled data (from bottom to top) and q = 1.5 (circle), 2.0 (square), and 2.5 (triangle). The straight lines are shown for comparison and have the slope γ for the long-term correlated data and 1 for the uncorrelated shuffled data.

see also: Altmann and Kantz 2005). Figure 19.2(b) shows a section of the sequence of the return intervals for Gaussian long-term correlated data for a threshold q (quantile) chosen such that the mean return interval R_q ('return period') is approximately 100. Figure 19.2(c) shows the same section, but the original data was shuffled before, destroying the correlations. One can see that there are more large r-values and many more short r-values in Fig. 19.2(b) compared to the uncorrelated case in Fig. 19.2(c), although the mean return interval R_q is the same. The long and short return intervals in (b) appear in clusters (Bunde et al. 2005), creating epochs of cumulated extreme events caused by the short r-values, and also long epochs of few extreme events caused by the long r-values. In the following we show, how R_q and the distribution density $P_q(r)$ of the return intervals are affected by the presence of long-term correlations as well as by different distribution densities $P(x)$ of the data.

Let us consider the mean return interval R_q. For a given threshold q, there exist N_q return intervals r_j, j = 1, 2, …, N_q, which satisfy the sum rule $\sum_{j=1}^{N_q} r_j = N$. When the data are shuffled, the long-term correlations are destroyed, but the sum rule still applies with the same value of N_q. Accordingly, for both long-term correlated and uncorrelated records, the mean return interval $R_q = N/N_q$ is not affected by the long-term correlations. This statement can also be considered as the time series analogous of Kac's Lemma (Kac 1947). Hence, R_q can be obtained directly from the tail of the (normalized) distribution density $P(x)$. Accordingly, there is a one-by-one correspondence between q and R_q, which depends only on the distribution density $P(x)$ but not on the correlations.

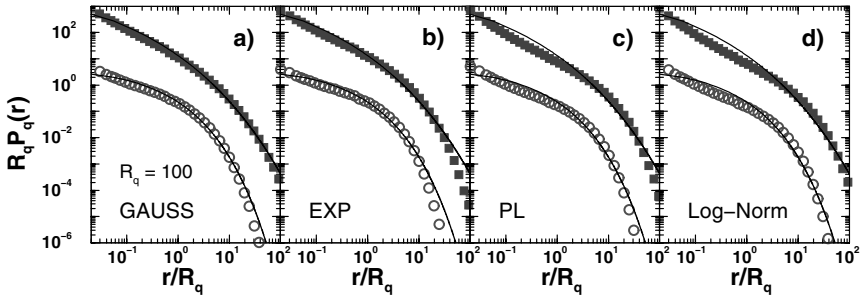


Fig. 19.4. Normalized rescaled distribution density functions $R_q P_q(r)$ of r-values with $R_q = 100$ as a function of r/R_q for long-term correlated data with $\gamma = 0.4$ (open symbols) and $\gamma = 0.2$ (filled symbols; we multiplied the data for the filled symbols by a factor 100 to avoid overlapping curves). In (a) the original data was Gaussian distributed, in (b) exponentially distributed, in (c) power-law distributed with power –5.5, and in (d) log-normally distributed. All four figures follow quite well stretched exponential curves (solid lines) over several decades. For small r/R_q values a power-law regime seems to dominate, while on large scales deviations from the stretched exponential behavior are due to finite-size effects.

2.2 Stretched exponential behavior and finite-size effects for large return intervals

It is known that for uncorrelated records ('white noise'), the return intervals are also uncorrelated and (according to the Poisson statistics) exponentially distributed with $P_q(r) = \frac{1}{R_q}\exp(-r/R_q)$, where R_q is the mean return interval for the given threshold q (see, e.g., (Storch and Zwiers 2001)). When introducing long-term correlations in Gaussian data the shape of $P_q(r)$ for large values of r ($r > R_q$) is changed to a stretched exponential (Bunde et al. 2003, Bunde et al. 2004, Bunde et al. 2005, Altmann et al. 2005, Eichner et al. 2007)

$$P_q(r) \cong \frac{a_\gamma}{R_q}\exp\left[-b_\gamma\left(r/R_q\right)^\gamma\right], \tag{3}$$

where the exponent γ is the correlation exponent, and the parameters a_γ and b_γ are independent of q. Their dependence upon γ can be determined from two normalization conditions. Hence, if γ is determined independently by correlation analysis, we can obtain a data collapse of all curves for different values of q by plotting $R_q P_q(r)$ versus r/R_q (according to Eq. 3) (Bunde et al. 2003, Bunde et al. 2004, Bunde et al. 2005). However, we have to note that Eq. 3 does not hold for small values of r, causing some deviations in the values of parameters a_γ and b_γ and the data collapse.

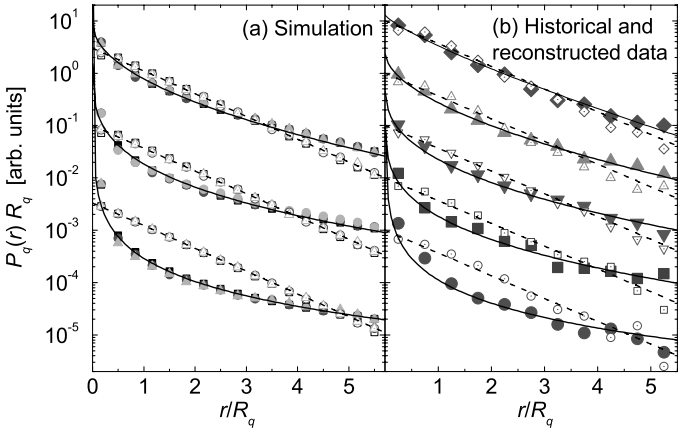


Fig. 19.5. (a) Distribution densities $P_q(r)$ of the return intervals r for the thresholds q = 1.5 ($R_q \approx 15$, squares), 2.0 ($R_q \approx 44$, circles), and 2.5 ($R_q \approx 161$, triangles) for simulated long-term correlated records with $\gamma = 0.5$ (top), 0.3 (middle), and 0.1 (bottom) (filled symbols) and for the corresponding shuffled data (open symbols). For the simulations, we used 1000 records of length $N = 2\ 10^6$ for each value of γ. (b) Distribution densities $P_q(r)$ of the return intervals r for the five climate records considered in Fig. 1(g) with same symbols, for both original data (filled symbols) and shuffled data (open symbols). The data have been averaged over all quantiles q with $R_q > 3$ y and more than 50 return intervals. The lines are the theoretical curves following Eq. 3. From Bunde et al. 2005.

Figure 19.3(a) displays the distribution densities $P_q(r)$ of the return intervals for long-term correlated ($\gamma = 0.4$) and uncorrelated Gaussian data, for 3 values of R_q, R_q = 15, 44, and 161. Figure 19.3(b) shows that there exists a data collapse for these 3 curves, when the axis are properly scaled with the return period R_q . This scaling feature is very important since it allows to make predictions also for large thresholds q where the statistics is poor. The continous line in the figure is stretched exponential with $\gamma = 0.4$, as suggested by Eq. (3). Figure 19.3(c) further emphasized the stretched exponential behavior. In the double-logarithmic plot of $-\ln(P_q(r)/P_q(1))$ versus r/R_q, the slopes of the lines are identical to the correlation exponents γ.

Figure 19.4(a-d) shows the rescaled distribution density function of the return intervals for one value of R_q, R_q = 100, for the four different types of distributions of the original data (Gaussian, exponential, power-law, and log-normal according to Eq. (2)) with correlation exponents $\gamma = 0.2$ and 0.4 in a double-logarithmic scale. The shape of the corresponding stretched exponentials, Eq. (3), is also plotted. The agreement is best for the Gaussian data, but also for the other distributions the stretched exponential fit is a good approximation over several decades.

Figure 19.5 compares $P_q(r)$ for simulated Gaussian data and three different correlation exponents with the five historical and reconstructed data sets introduced in Fig. 19.1. The solid lines, representing the theoretical curves with the measured γ values, match with $P_q(r)$ of the data (filled symbols). The dotted lines and the open symbols show the results for the shuffled data, when all correlations are destroyed. The shape of $P_q(r)$ becomes a simple exponential.

2.3 Power-law regime and discretization effects for small return intervals

The curves in Fig. 19.4 exhibit significant deviations from the stretched exponential form also for small values of r ($r < R_q$) which we have studied in detail. It reveals the occurrence of an additional intermediate scaling regime for $r < R_q$. The scaling behavior with R_q still holds in this new regime, such that data for different values of q collapse onto each other as long as $r >> 1$. The scaling behavior in this regime might be characterized by a power-law, giving rise to the two-branched ansatz

$$R_q P_q(r) \sim \begin{cases} \left(r/R_q\right)^{\gamma'-1} & \text{for } 1 << r < R_q \\ \exp\left[-b_\gamma \left(r/R_q\right)^{\gamma}\right] & \text{for } R_q < r << N, \end{cases} \tag{5}$$

replacing Eq. (3). For all distribution densities $P(x)$, $\gamma' \approx \gamma$ seems to be consistent with the data. However, we cannot fully exclude that γ' might depend slightly on the quantile q.

The behavior of $P_q(r)$ as described by Eq. 3 and the large r branch of Eq. 5 becomes visible only for very long data. However, for short natural records or large values of q and R_q, where the asymptotic regime $R_q < r << N$ can hardly be observed because N does not significantly exceed R_q the power-law regime will clearly dominate the scaling behavior of the return interval distribution density. Hence, power-law distributions of return intervals, for example, observed in a recent analysis of waiting times (Oliveira and Barabasi 2005; Vazquez, Oliveira, Dezsö, Goh, Kondor, and Barabasi 2006) might be considered as a possible indication of long-term correlations.

For very small r values (close to 1) the continuous distribution density Eq. 5 has to be replaced by a discrete distribution. Hence, the data points for r close to 1 cannot obey the scaling of the distribution density with R_q, and no data collapse can be achieved for them. For the power-law distributed data and the log-normal distributed data, the discretization effects even seem to suppress the full development of the power-law scaling regime.

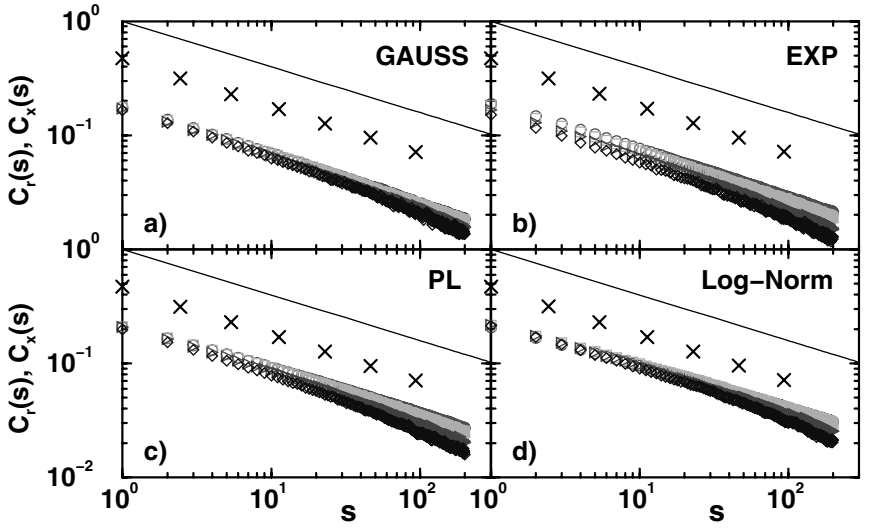


Fig. 19.6. Autocorrelation function $C_x(s)$ of the original data (cross) and $C_r(s)$ (see Eq. 6) of series of return intervals: (a) Gaussian, (b) exponentially, (c) power-law, and (d) log-normally distributed original data with $\gamma = 0.4$ for $R_q = 10$ (circle), 20 (square), 50 (triangle), and 100 (diamonds). All curves show a power-law behavior indicating long-term correlations in the sequence of values with the same correlation exponent given in the original data. Curves for large R_q show weaker correlations due to finite-size effects. All results were averaged over 150 configurations of original data sets of length $N = 2^{21}$. Solid lines are guides to the eye with slopes $-\gamma = -0.4$. After Eichner et al. 2007.

2.4 Long-term correlations of the return intervals

The form of the distribution density $P_q(r)$ of return intervals between extreme events in long-term correlated data indicates that very short and very long return intervals r are more frequent than for uncorrelated data. However, $P_q(r)$ does not quantify, if the return intervals themselves are arranged in a correlated fashion, and if clustering of rare events may be induced by long-term correlations. In our previous work (Bunde et al. 2004, Bunde et al. 2005) we reported that (i) long-term correlations in a Gaussian time series induce long-term correlations in the sequence of return intervals and (ii) that both correlation functions are characterized by the same correlation exponent γ. We showed that this leads to a clustering of extreme events, an effect that also can be seen in long climate records.

In order to determine the autocorrelation behavior of the return series r_j , $j = 1, 2 ...$, N_q for a given quantile q we calculate the autocorrelation function

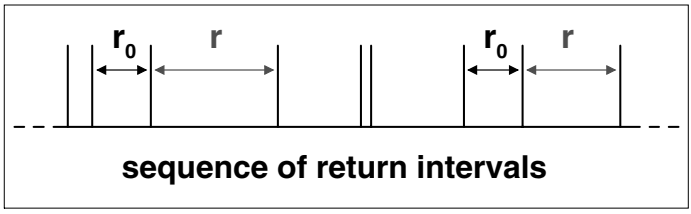


Fig. 19.7. Conditional return intervals are those r values, that follow in the sequence of the return intervals directly after a r value of a distinct size r_0. For data with poor statistics, it is convenient to use r_0 as a boundary value, i. e., considering all those r values that follow directly after a r value smaller or larger than, e. g., the median of all return intervals.

$$C_r(s) = \frac{1}{\sigma_r^2 (N-s)} \sum_{j=1}^{N_q - s} (r_j - R_q)(r_{j+s} - R_q). \tag{6}$$

Figure 19.6 shows $C_r(s)$ and $C_x(s)$ for data characterized by the four distribution densities *D(x)* and four return periods R_q. One can see that for each distribution, the data collapse to a single line, exhibiting the same slope as the original data. This shows that the return intervals are also long-term correlated, with the same value of γ as the original data. There is, however, one important difference in the correlation behavior: For the return intervals, the autocorrelation function $C_r(s)$ is significantly below the autocorrelation function $C_x(s)$ of the original data (also shown in Fig.19.6) by a factor between 2 and 3, depending on the distribution. Accordingly, there is additional white noise in the return interval sequences that does not depend on the return period R_q. We believe that this uncorrelated component is a consequence of the way the return intervals are constructed. Tiny changes in the threshold q can lead to large changes of several return intervals in the records, and this causes the additional random component. The strength of this random component is independent of the threshold q as confirmed by the scaling of the autocorrelation function $C_r(s)$ for different R_q . There is no crossover in the scaling behavior of the autocorrelation function except for finite-size effects.

Alternatively to an autocorrelation analysis via autocorrelation function one might want to employ the detrended fluctuation analysis (DFA, see introduction) and use the relation $\alpha = 1 - \gamma/2$ between the fluctuation exponent α and the correlation exponent γ to determine γ (Kantelhardt et al. 2001). In this case, however, a crossover is observed in the scaling behavior of the fluctuation function $F(s)$ because the uncorrelated component affects $F(s)$ also for $s > 1$. Hence, the classical autocorrelation analysis is more

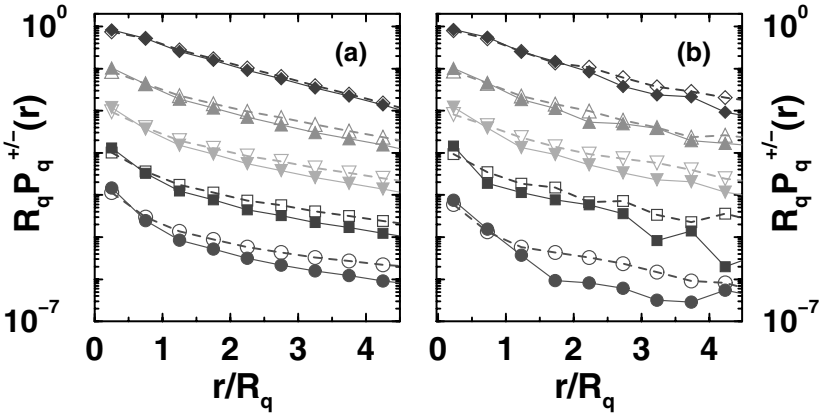


Fig. 19.8. Conditional distribution densities $R_q\ P_q^+(r)$ (open symbols) and $R_q\ P_q^-(r)$ (full symbols), averaged over all r_0 above and below the median return interval, respectively, versus r/R_q for (a) artificial records and (b) the five climate records from Fig. 1. The artificial data in (b) have the same γ values and mean record lengths as the climate records; we studied 1000 records of size $N = 1250$. After Bunde et al. 2005.

suitable for the study of the correlation properties of the return interval series than sophisticated DFA approaches.

2.5 Conditional Return Intervals

The long-term correlations in the sequence of the return interval series studied in the previous Section cause a dependence of the probability of finding a certain return interval r on the history. In order to study this effect we considered the *conditional* distribution density $P_q(r|r_0)$, that is the distribution density of all those return intervals that directly follow a return interval of given size r_0. An explanation of conditional return intervals is shown in Fig. 19.7. For fixed r_0/R_q the data points still collapse to single curves, independent of the distribution of the data. Taking Eq. 5 into account our results show that the conditional distribution density function scales as

$$P_q(r \mid r_0) \cong \frac{1}{R_q} f_{r_0/R_q}\left(r/R_q\right). \tag{7}$$

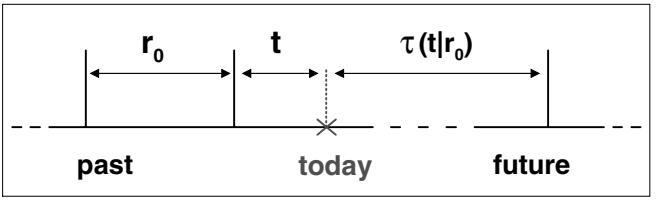


Fig. 19.9. The conditional residual waiting time $\tau_q(t|r_0)$ is defined as the average remaining time between today and the next extreme event, when the time between today and the last extreme event is t and the return interval between the two last extreme events is r_0.

The long-term correlations in the sequence of r values cause a culmination of small r values for small r_0 and large r values for large r_0. The conditional distribution densities for the five historical and reconstructed records and for artificial data with the same correlation exponents and similar set lengths are show in Fig. 19.8 for just two conditions, r_0 smaller (–) or larger (+) the median r value. The splitting of the curves $P_q^+(r)$ and $P_q^-(r)$ for large r/R_q is a consequence of the long-term correlations.

2.6 Conditional mean return intervals

In contrast to R_q the conditional mean return interval $R_q(r_0) = \sum_{r=1}^{\infty} rP_q(r\,|\,r_0)$, i. e. the average return interval of those r values that follow directly after an interval of size r_0, clearly exhibits correlation effects. Figure 19.10 shows $R_q(r_0)$ in units of R_q as a function of r_0/R_q for four values of R_q for long-term correlated data ($\gamma = 0.4$) following the four distribution densities $P(x)$ listed in Eq. 2. The correlation effect becomes apparent: after small r_0/R_q the next expected return interval $R_q(r_0)$ is smaller than R_q, and after large r_0/R_q it is much larger than R_q. Although the shapes of the (more or less) collapsing curves differ depending on the original distribution and on the chosen quantile q (i.e. on R_q), the tendency is the same for all four original distribution densities.

By definition, $R_q(r_0)$ is the expected waiting time to the next event, when the two events before were separated by r_0. A more general quantity is the expected residual waiting time $\tau_q(t|r_0)$ to the next event, when the time t has been elapsed (see Fig. 19.9). For $t = 0$, $\tau_q(t|r_0)$ is identical to $R_q(r_0)$. In general, $\tau_q(t|r_0)$ is related to $P_q(r|r_0)$ by

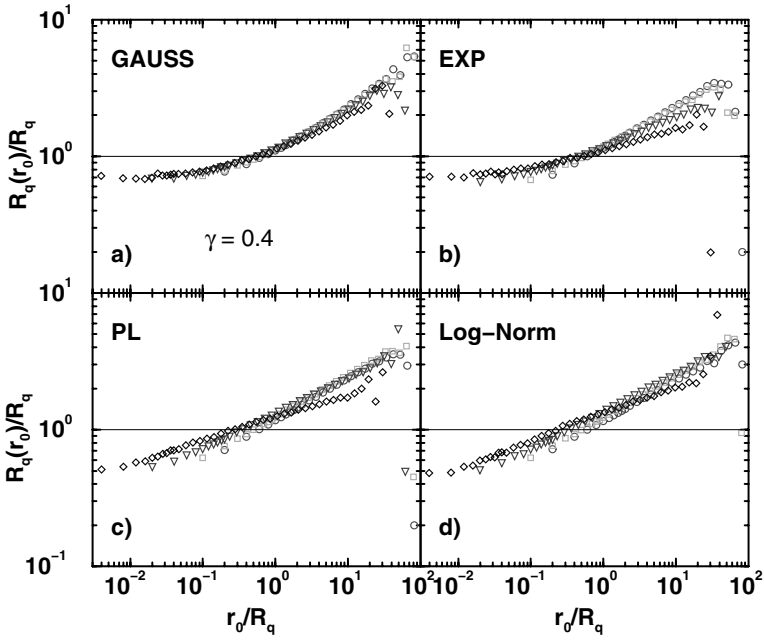


Fig. 19.10. Conditional return periods $R_q(r_0)$ in units of R_q versus the condition r_0/R_q for R_q = 5 (circles), 10 (squares), 50 (triangles), and 250 (diamonds) for $\gamma = 0.4$. While Gaussian (a) and exponentially (b) distributed data show a data collapse with nearly no R_q-dependence (except for stronger finite-size effects for exponential data at large r_0/R_q), power-law (c) and log-normally distributed (d) data show slightly different slopes for different R_q. All figures clearly display the memory-effect in form of increasing $R_q(r_0)$ with increasing r_0. In uncorrelated data $R_q(r_0) = R_q$, as indicated by the horizontal lines at ratio one. After Eichner et al. 2007.

$$\tau_q(t \mid r_0) = \int_t^\infty (r-t) P_q(r \mid r_0)\, dr \Big/ \int_t^\infty P_q(r \mid r_0)\, dr\,. \tag{8}$$

For uncorrelated records, $\tau_q(t|r_0) = R_q$ (except for discreteness effects that lead to $\tau_q(t|r_0)/R_q > 1$ for $t > 0$, see (Sornette and Knopoff 1997). Due to the scaling of $P_q(r|r_0)$, we expect that also $\tau_q(t|r_0)/R_q$ scales with r_0/R_q and t/R_q. Figure 19.11(a) shows that this is indeed the case. The data collapse for each value of t/R_q confirms the scaling property. The figure clearly displays the effect of the long-term memory: Small and large return intervals are more likely to be followed by small and large ones, respectively, and hence $\tau_q(t|r_0)/R_q \equiv R_q(r_0)/R_q$ is well below (above) 1 for r_0/R_q well below (above) 1. With increasing t, the expected residual time to the next event increases, as is also shown in Fig. 19.11(b), for two values of r_0 (top and bottom curve). Note that only for an infinite long-term correlated record, the value of $\tau_q(t|r_0)$ will

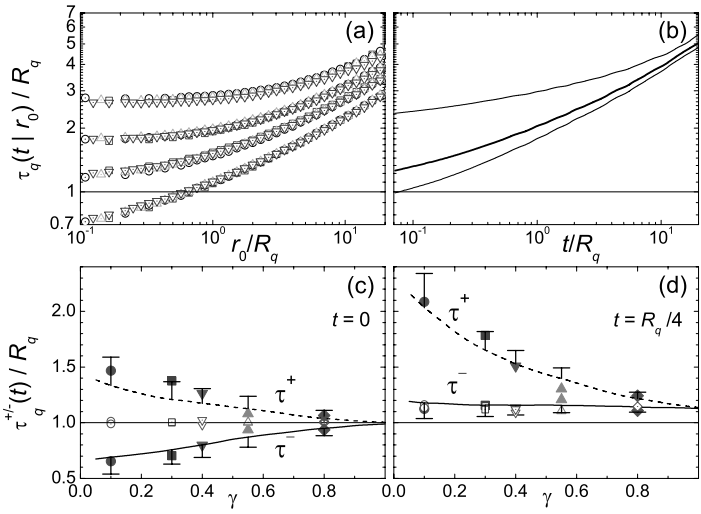


Fig. 19.11. (a) Mean conditional residual time to the next event $\tau_q(t|r_0)$ (in units of R_q) versus r_0/R_q for four q values (q = 1.0, 1.25, 1.5 and 1.75, different symbols) and four values of the elapsed time t since the last event (t/R_q = 0, 0.25, 1 and 4, from bottom to top). (b) Mean residual time $\tau_q(t|r_0)$ as a function of t/R_q, for r_0/R_q = 1/8 (lower curve) and r_0/R_q = 8 (upper curve). The middle curve represents the mean conditional residual time averaged over all r_0. (c,d) Mean conditional residual times $\tau_q^-(t)$ (full line) and $\tau_q^+(t)$ (dashed line), averaged over all r_0 below and above the median return interval, respectively, for (c) t = 0 and (d) $t = R_q/4$. The symbols are for the five climate records from Fig. 1(g) and Fig. 5(b), for both original data (filled symbols) and shuffled data (open symbols). After Bunde et al. 2005.

increase indefinitely with t and r_0. For real (finite) records, there exists a maximum return interval which limits the values of t, r_0 and $\tau_q(t|r_0)$. The middle curve shows the expected residual time averaged over all r_0, i.e. the unconditional residual time. In this case, the interval between the last two events is not taken explicitly into account, and the slower-than-Poisson-decrease of the unconditional distribution density $P_q(r)$, leads to the anomalous increase of the mean residual waiting time with the elapsed time (Sornette et al. 1997). Very recently, this approach (average over r_0) has been applied successfully to worldwide earthquake records (Corral 2004). For the case of long-term correlated records, however, like the hydroclimate records discussed here, the large differences between the three curves in Fig. 19.11(b) suggest

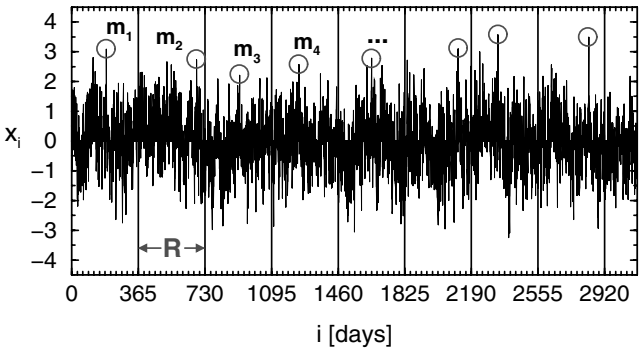


Fig. 19.12. Definition of maxima: A time series (x_i), $i = 1, \ldots, N$, of, e. g., daily data is separated into segments of length $R = 365$ days. The maximum values m_j (circles) in each segment, e. g., annual maxima, define another time series (m_j), $_j = 1, \ldots, N/R$. From Eichner et al. 2006.

that for an efficient risk estimation, also the previous return interval has to be taken into account, and not only the distribution of the return intervals.

To reveal this intriguing behavior in the relatively short observed and reconstructed records, we improved the statistics (similar to Fig. 19.8) by studying the mean residual waiting times $\tau_q^-(t)$ and $\tau_q^+(t)$ for r_0 below and above the median, respectively. For uncorrelated data, both quantities are identical and coincide with R_q.

Figure 19.11(c) shows $\tau_q^-(0)/R_q$ and $\tau_q^+(0)/R_q$ versus γ for simulated records (lines) and the five representative climate records (symbols). The difference between $\tau_q^-(t)$ and $\tau_q^+(t)$ becomes more pronounced with decreasing value of γ, i. e. increasing long-term memory. The results for the climate records are in good agreement with the theoretical curves. The same comparison for $t/R_q = 1/4$ instead of $t = 0$ is shown in Fig. 19.11(d). The behaviour is qualitatively different: while $\tau_q^-(0)/R_q$ increases with increasing γ, $\tau_q^-(R_q/4)/R_q$ is rather constant. Again, the agreement between simulated and real records is quite satisfactory, revealing the strong effect of memory in the hydroclimate records that also results in the clustering of the extreme events. To show the significance of the results, we also analyzed the corresponding shuffled data. We obtained $\tau_q^+(0)/R_q \approx \tau_q^-(0)/R_q \approx 1$ and $\tau_q^+(R_q/4)/R_q \approx \tau_q^-(R_q/4)/R_q \approx 1.1$. In the second case, the shuffled data (following the Poisson distribution) show a slight increase of the residual time (1.1 instead of 1). This is a finite-size effect that already has been noticed in (Sornette et al. 1997).

3 Statistics of Maxima

Extreme events are rare occurrences of extraordinary nature, such as floods, very high temperatures, or major earthquakes. In studying the extreme value statistics of the corresponding time series one wants to learn about the distribution of the extreme events, i. e., the maximum values of the signal within time segments of fixed duration R, and the statistical properties of their sequences. Figure 19.12 illustrates the definition of the series of maxima (m_j), $_j = 1, \ldots, N/R$ of original data (x_i), $i = 1, \ldots, N$, within segments of size R for $R = 365$, i. e. for annual maxima if (x_i) represents daily data.

Many exact and empirical results on extreme value statistics have been obtained in the past years, for reviews see, e. g., (Gumbel 1958; Galambos 1978; Leadbetter, Lindgren and Rootzen 1983; Galambos, Lechner and Simin 1994; Embrechts, Klüppelberg, and T. Mikosch 1997; v. Storch and Zwiers 2001). Most of these results, however, hold only in the limit $R \to \infty$ and are based on statistically independent values of the time series. Both assumptions are not strictly valid in practice. Since observational data are always finite, predictions for finite time intervals R are required, and – most importantly – correlations cannot be disregarded (Eichner, Bunde, Kantelhardt and Havlin 2006).

3.1 Extreme value statistics for i.i.d. data

In classical extreme value statistics one assumes that records (x_i) consist of independently and identically distributed (i.i.d.) data, described by density distributions $P(x)$, which can be, e. g., a Gaussian or an exponential distribution. One is interested in the distribution density function $P_R(m)$ of the maxima (m_j) determined in segments of length R in the original series (x_i) (see Fig. 19.12). Note that all maxima are also elements of the original data. The corresponding integrated maxima distribution $G_R(m)$ is defined as

$$G_R(m) = 1 - E_R(m) = \int_{-\infty}^{m} P_R(m')\, dm\,. \tag{9}$$

Since $G_R(m)$ is the probability of finding a maximum smaller than m, $E_R(m)$ denotes the probability of finding a maximum that exceeds m. One of the main results of traditional extreme value statistics states that the integrated distribution $G_R(m)$ converges to a double exponential (Fisher-Tippet-Gumbel) distribution (often labeled as Type I) (Fisher and Tippett 1928; Gumbel 1958; Galambos 1978; Leadbetter et al. 1983) for i.i.d. data with Gaussian or exponential $P(x)$, i. e.,

$$G_R(m) \to G\left(\frac{m-u}{\alpha}\right) = \exp\left[-\exp\left(-\frac{m-u}{\alpha}\right)\right] \tag{10}$$

for $R \to \infty$, where α is the scale parameter and u the location parameter. By the method of moments the parameters are given by $\alpha = \sigma_R\sqrt{6}/\pi$ and $u = m_R - n_e\alpha$ with the Euler constant $n_e = 0.577216$ (Leadbetter et al. 1983; Chow 1964; Raudkivi

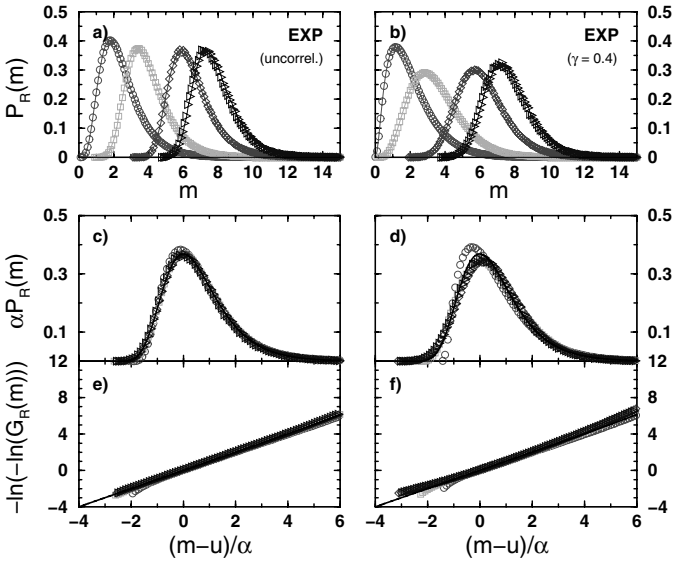


Fig. 19.13. Distributions of maxima in segments of length R for (a,c,e) uncorrelated and (b,d,f) long-term correlated ($\gamma = 0.4$) data with exponential distribution. Pannels (a,b) show the distribution density function $P_R(m)$ of the maximum values for four segment sizes $R = 6$ (circles), 30 (squares), 365 (diamonds), and 1500 (triangles). Pannels (c,d) show that a collapse of all four curves to a single curve is achieved in both cases, when the m axis is replaced by $(m - u)/\alpha$ and $P_R(m)$ is multiplied by the scale parameter α. The solid line is the Gumbel distribution density, Eq. 11. Pannels (e,f) show the corresponding integrated distribution $G_R(m)$ together with the Gumbel function Eq. 10. After Eichner et al. 2006.

1979; Rasmussen and Gautam 2003). Here m_R and σ_R denote the (R dependent) mean maximum and the standard deviation, respectively. Note that different asymptotics will be reached for broader distributions of data (x_i) that belong to other domains of attraction (Leadbetter et al. 1983). For example, for data following a power-law (or Pareto) distribution, $P(x) = (x/x_0)^{-k}$, $G_R(m)$ converges to a Fréchet distribution, often labeled as Type II. For data following a distribution with finite upper endpoint, for example the uniform distribution $P(x) = 1$ for $0 \le x \le 1$, $G_R(m)$ converges to a Weibull distribution, often labeled as Type III. These are the other two types of asymptotics, which, however, we do not consider here.

3.2 Effect of long-term persistence on the distribution of the maxima

In long-term correlated records extreme events cannot be viewed a priori as uncorrelated even when there is a long time span between them. Recently, there have

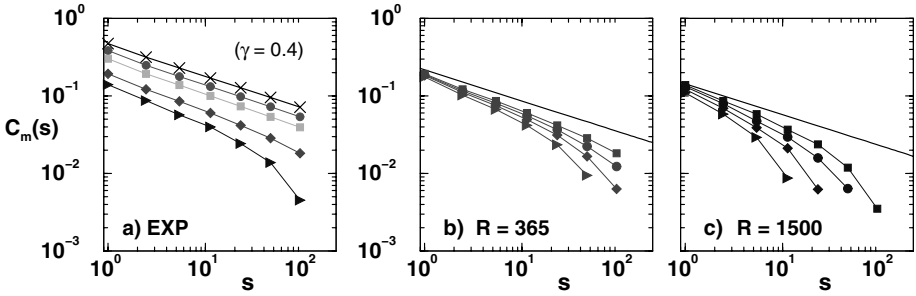


Fig. 19.14. (a) Autocorrelation function $C_m(s)$ of the maxima (m_j) of Gaussian distributed (x_i) for different R values, R = 6 (circles), 30 (squares), 365 (diamonds), and 1500 (triangles). The autocorrelation function $C_x(s)$ of the original data (x_i) (crosses) shows the slope $-\gamma = -0.4$. (b,c) Study of finite-size effects with (b) R = 365 and (c) 1500. The set lengths are $N = 2^{21}$ (squares), 2^{20} (circles), 2^{19} (diamonds), and 2^{18} (triangles). The descent of the slopes of $C_m(s)$ from the slope of the straight line ($-\gamma = -0.4$) with decreasing set length seems to be a finite-size effect. After Eichner et al. 2006.

been some approaches to include correlations in the study of extreme value statistics. For the special case of Gaussian 1/*f* correlations in voltage fluctuations in GaAs films extreme value statistics have been demonstrated to follow a Gumbel distribution (Antal, Droz, Györgyi, and Racz 2001). A somewhat different asymptotic behavior was observed in experiments on turbulence and in the two-dimensional XY model (Bramwell, Holdsworth and Pinton 1998; Bramwell, Christensen, Fortin, Holdsworth, Jensen, Lise, Lopez, Nicodemi, Pinton, and Sellitto 2000), see also (Dahlstedt and Jensen 2001). Extreme value statistics have also been employed in studies of hierarchically correlated random variables representing the energies of directed polymers (Dean and Majumdar 2001) and in studies of maximal heights of growing self-affine surfaces (Raychaudhuri, Cranston, Przybyla and Shapir 2001). In the Edwards-Wilkinson model and the Kardar-Parisi-Zhang model for fluctuating, strongly correlated interfaces an Airy distribution function has been obtained as exact solution for the distribution of maximal heights very recently (Majumdar and Comtet 2004/2005). On the other hand, the statistics of extreme height fluctuations for Edwards-Wilkinson relaxation on small-world substrates are rather described by the classical Fisher-Tippet-Gumbel distribution (Guclu and Korniss 2004). Besides these recent results there is a theorem by S. M. Berman (Berman 1964) (see also Galambos 1978; Leadbetter et al. 1983) stating that the maxima statistics of stationary Gaussian sequences with correlations converges to a Gumbel distribution asymptotically for $R \to \infty$ provided that $C_x(s)\backslash\log(s) \to 0$ for $s \to \infty$, which holds for long-term correlations.

Figure 19.13 shows how the convergence of the integrated maxima distribution $G_R(m)$ towards the Gumbel distribution, Eq. 10, is affected by long-term correlations in exponentially distibuted signals (x_i). In pannels (a,b) the unscaled distribution densities $P_R(m)$ of the maxima within segments of size R are shown for several values of R. Since Eqs. 9 and 10 yield that for $R \to \infty$

$$P_R(m) \to \frac{1}{\alpha}\exp\left[-\exp\left(-\frac{m-u}{\alpha}\right)-\frac{m-u}{\alpha}\right], \tag{11}$$

the distribution densities $P_R(m)$ can be scaled upon each other if $\alpha P_R(m)$ is plotted versus $(m - u)/\alpha$, see Fig. 19.13(c,d). In Fig. 19.13(c,d) deviations occur only at very small R values ($R < 10$) where scaling breaks down due to the sharp cutoff of the exponential density distribution $P(x)$ at $x = 0$. In the long-term correlated case, where the correlation time T diverges, the fast convergence is particularly surprising, since the segment duration R can never exceed T. From a theoretical point of view, we expect a convergence towards the Gumbel limit only for very large R values. The reason for this fast convergence may be a rapid weakening of the correlations among the maxima with increasing values of R.

In Fig. 19.13(e,f) it is shown that the convergence towards Eq. 10 (continuous line) is also fast for the exponentially distributed data. In contrast, the convergence is much slower in the case of a Gaussian distribution of the original data (Eichner et al. 2006). Hence, the distribution $P(x)$ of the original data has a much stronger effect upon the convergence towards the Gumbel distribution than the long-term correlations in the data.

3.3 Effect of long-term persistence on the correlations of the maxima

The distributions of maxima considered in the previous Section do not quantify, however, if the maxima values are arranged in a correlated or in an uncorrelated fashion, and if clustering of maxima may be induced by long-term correlations in the data. To study this question, we have evaluated the correlation properties of the series of maxima (m_j), $j = 1, \ldots, N/R$, of long-term correlated data with Gaussian distribution. Figure 19.14(a) shows representative results for the maxima autocorrelation function

$$C_m(s) = \frac{\left\langle (m_j - m_R)(m_{j+s} - m_R)\right\rangle}{\left\langle (m_j - m_R)^2\right\rangle}, \tag{12}$$

where m_R denotes the average maximum value in the series, and $\langle \ldots \rangle$ is the average over j similar to Eq. 1. The comparison with the scaling behavior of the auto-correlation function $C_x(s)$ of the original data (x_i) (see Eq. 1) that follows a power law decay, $C_x(s) \sim s^{-\gamma}$ with $\gamma = 0.4$, reveals the presence of long-term correlations with a correlation exponent $\gamma' \approx \gamma$ in the maxima series. Hence, large maxima m are more likely to be followed by large maxima and small maxima are rather followed by small maxima, leading to clusters of large and small maxima.

Figures 19.14(b,c) show that the deviations of the auto-correlation function $C_m(s)$ from a power-law fit with slope $\gamma = -0.4$ for large values of R and s are presumably caused by finite-size effects. They become significantly smaller as the length N of

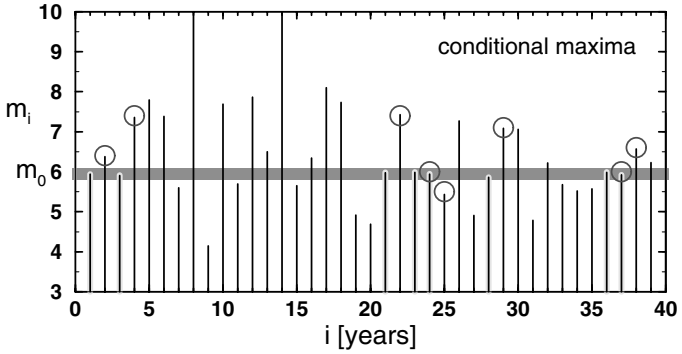


Fig. 19.15. Definition of conditional maxima. In the sequence of (annual) maxima only those m values (indicated by circles) are considered, which directly follow a maximum of approximate size $m_0 \approx 6$ (gray band). The new sequence of m values is the sequence of conditional maxima. From Eichner et al. 2006.

the series is increased. In the case of uncorrelated data, the series of maxima is also uncorrelated, $C_m(s) = 0$ for $s > 0$ (not shown).

3.4 Conditional mean maxima

As a consequence of the long-term correlations in the series of maxima (m_j), the probability of finding a certain value m_j depends on the history, and in particular on the value of the immediately preceeding maximum m_{j-1}, which we will denote by m_0 in the following. This effect has to be taken into account in predictions and risk estimations. For a quantitative analysis we considered conditional maxima as illustrated in Fig. 19.15, where all maxima following an $m_0 \approx 6$ (within the grey band), i. e., the subset of maxima which fulfill the condition of having a preceeding maximum close to m_0, are indicated by circles. The width Δm_0 sketched by the grey band around m_0 in Fig. 19.15 is set such that a sufficient number of approximately 700 conditional maxima is obtained for each record. The corresponding conditional mean maximum value $m_R(m_0)$ is defined as the average of all these conditional maxima. Note that $m_R(m_0)$ will be independent of m_0 for uncorrelated data.

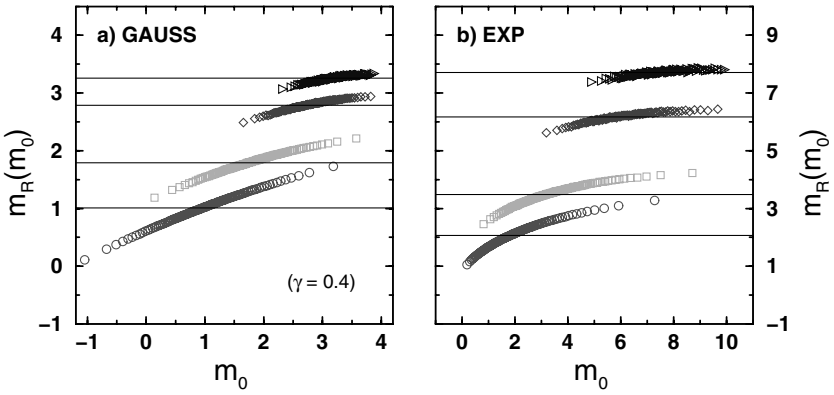


Fig. 19.16. Conditional mean maxima $m_R(m_0)$ for $\gamma = 0.4$ and $R = 6$ (circles), 30 (boxes), 365 (diamonds), and 1500 (triangles) versus m_0 for(a) Gaussian and (b) exponential data. The straight lines represent the unconditional means m_R for given R. The width Δm_0 for the condition m_0 was chosen such that approximately 700 m values were obtained for each m_0 in each of the 150 runs of $N = 2^{21}$ x values. Both figures show the memory effect in form of $m_R(m_0) > m_R$ for rather large m_0 (above m_R) and $m_R(m_0) < m_R$ for rather small m_0 (below m_R). From Eichner et al. 2006.

Figure 19.16 shows the conditional mean maxima $m_R(m_0)$ versus m_0 for long-term correlated Gaussian and exponentially distributed data for four values of R. Of course, the mean maxima are larger for larger segment sizes R. This dependence is also observed for the unconditional mean maxima indicated by horizontal lines in Fig. 19.15. In addition to this trivial dependence, the conditional mean maxima significantly depend upon the condition, i. e. the previous maximum m_0, showing a clear memory effect. Evidently, this dependence is most pronounced for the small segment durations. However, it is still observable for the large $R = 365$ (most common for observational daily data) and even $R = 1500$ (beyond common observational limits). Note that the results for Gaussian and exponentially distributed data agree only qualitatively: while the m_0 dependence of $m_R(m_0)$ is quite close to linear for Gaussian data, there seems to be significant curvature for the exponentially distributed data, which is a remnant of the asymmetry of the exponential distribution. To test our predictions on real records with long-term correlations we studied the annual data of the Nile river water level minima (Whitcher et al. 2002; see also Fig. 19.1) and the reconstructed northern hemisphere annual temperatures by Moberg (Moberg, Sonechkin, Holmgren, Datsenko, and Karlén 2005). Since the Nile data consists of annual minima, we studied extreme minima instead of maxima. Both records are long-term correlated with $\gamma \approx 0.3$ in Eq. 1.

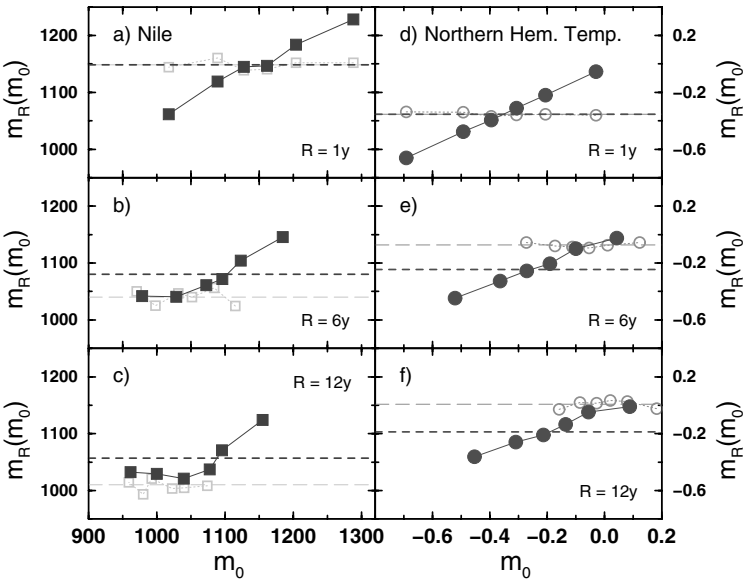


Fig. 19.17. (a-c) Conditional mean minima $m_R(m_0)$ for annual data of Nile river water level minima (squares) (Whitcher et al. 2002) and (d-f) conditional mean maxima for the reconstructed northern hemisphere annual temperatures after Moberg (circles) (Moberg et al. 2005), for (a,d) R = 1y, (b,e) 6y, and (c,f) 12y. The filled symbols show the results for the real data and the open symbols correspond to surrogate data where all correlations have been destroyed by random shuffling. The unconditional mean minima (a-c) and maxima (d-f) are indicated by dashed lines; the long-dashed lines correspond to the shuffled data. From Eichner et al. 2006.

In order to get sufficient statistics for the conditional means $m_R(m_0)$, we have considered six m_0 intervals for each value of R and have set the width Δm_0 of the band around m_0 such that there are no gaps between the bands. Figure 19.17 shows the results for three values of R. In all cases, the effect of the long-term correlations (persistence) on the conditional mean minima and maxima $m_R(m_0)$is clearly visible for both records: the conditional means are smaller for smaller condition value m_0 and larger for larger condition value. As shown by the open symbols in Fig. 19.17 the m_0 dependence completely disappears for randomly shuffled surrogate data, indicating that the dependence was due to the correlations in the data.

3.5 Conditional maxima distributions

The quantity $m_R(m_0)$ is the first moment of the conditional distribution density $P_R(m|m_0)$, which is defined as the distribution density of all maxima m_j that follow a

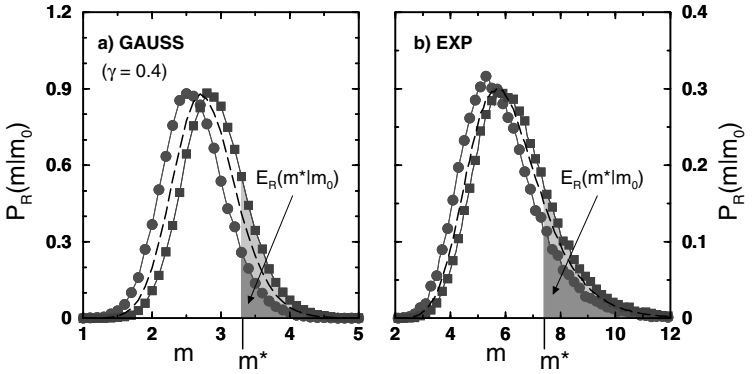


Fig. 19.18. (a) Conditional distribution density $P_R(m|m_0)$ of maximum values taken from correlated Gaussian data ($\gamma = 0.4$) with $R = 365$ and $m_0 = 2.06$ (circles) as well as $m_0 = 3.55$ (squares). Figure (b) shows the same as (a) for exponentially distributed data with $m_0 = 4.10$ (circles) and $m_0 = 8.65$ (squares). The probability $E_R(m^*|m_0)$ to find a m value larger than an arbitrarily given m^* (see Eq. 13) also depends on m_0. From Eichner et al. 2006.

given maximum value m_0. Figure 19.18 shows $P_R(m|m_0)$ for two values of m_0 and again for Gaussian as well as for exponentially distributed long-term correlated data sets with $\gamma = 0.4$. When compared with the unconditional distribution density $P_R(m)$, the long-term correlations lead to a shift of $P_R(m|m_0)$ to smaller m values for small m_0 and to larger m values for large m_0, respectively. The conditional exceedance probability

$$E_R(m \mid m_0) = \int_m^\infty P_R(m' \mid m_0)\, dm' \tag{13}$$

defines the probability of finding a maximum larger than m provided that the previous value was close to m_0. We found a strong dependence of $E_R(m|m_0)$ upon the condition m_0. Consequently, the difference between the unconditional probabilities $E_R(m)$ and the corresponding conditional probabilities $E_R(m|m_0)$ depends strongly on m_0 in the presence of long-term correlations.

Figure 19.19 shows the ratios of the conditional exceedance probabilities $E_R(m|m_0)$ and the unconditional exceedance probabilities $E_R(m)$. The figure clearly shows an increase of the memory effect for larger m values, i. e., for more extreme events. This increase seems weaker for exponentially distributed data than for Gaussian distributed data due to the less correlated maxium series of exponential data; however the tendency is the same. As Fig. 19.19 shows $E_R(m|m_0)$ can differ up to a factor of two from $E_R(m)$ when considering the history m_0 in the presence of long-term correlations (with $\gamma = 0.4$). This effect has to be taken into account in predictions and risk estimations of large events.

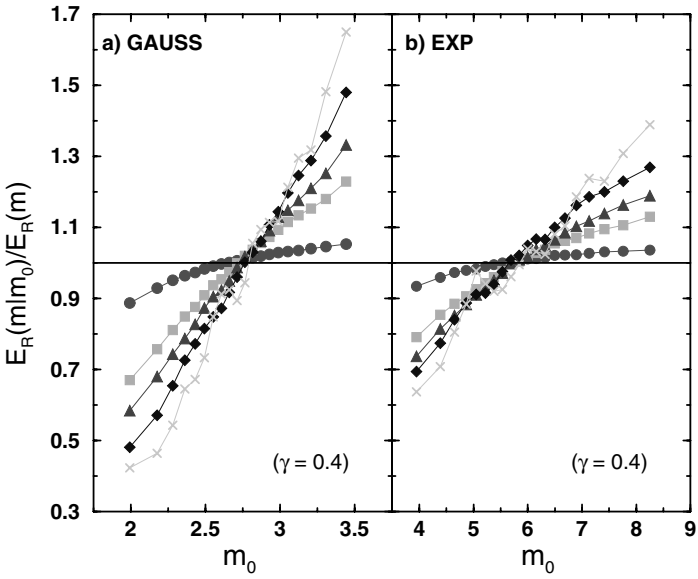


Fig. 19.19. Ratios of the conditional and unconditional exceedance probabilities $E_R(m|m_0)$ and $E_R(m)$ to find a maximum larger than m for (a) Gaussian and (b) exponentially distributed data with $\gamma = 0.4$ and $R = 365$. The six m values were chosen such that $E_R(m) = 0.9$ (circles, m = 2.15 for Gaussian and 4.40 for exponential data), 0.5 (squares, m = 2.75 and 5.95), 0.3 (triangles up, m = 2.95 and 6.70), 0.1 (diamonds, m = 3.35 and 8.00), and 0.01 (crosses, m = 3.95 and 10.40). Each point in the graph is based on astatistics of 500 conditional m values and averaged over 150 runs of $N = 2^{21}$ x values. The effect of long-term correlations seems to be strongest for the largest m (crosses): depending on m_0 $E_R(m|m_0)$ varies from 0.4 up to 1.7 for Gaussian data, i. e., by a factor greater than four. For exponential data this factor is still greater than two. From Eichner et al. 2006.

4 Conclusion

In summary, we have studied the effect of long-term correlations in time series upon extreme value statistics and the return intervals between extreme events. For the return intervals in long-term persistent data with Gaussian, exponential, power-law, and log-normal distribution densities, we have shown that mainly the correlations rather than the distributions affect the return interval statistics, in particular the distribution density of return intervals, the conditional distribution density of return intervals, the conditional mean return intervals, and the correlation properties of the return interval series. The stretched exponential decay of the return interval distribution density for long return intervals is complemented by a power-law decay for small return intervals, which will dominate the observable behavior in most

(rather short) observational data. Still, the scaling behavior with the mean return interval holds in both regimes. We have also shown that the long-term persistence inherent in hydroclimate records represents a natural mechanism for the clustering of the hazardous events. We also found that, as a consequence of the long-term memory, the mean residual waiting time to the next event increases with the elapsed time and depends strongly on the previous return interval. We have demonstrated that also this counterintuitive phenomenon can be seen in long climate records.

Considering series of maxima within segments of size R of the original data, we have shown numerically that the maxima distribution functions still converge to the same type of Gumbel distributions as for uncorrelated data for increasing R. For finite values of R, however, some deviations occur especially for originally Gaussian distributed data. Our extensive numerical simulations suggested that contrary to the common assumption in extreme value statistics, the maxima time series turn out to be *not* independently, identically distributed numbers. The series of maxima rather exhibit long-term correlations similar to those in the original data. Most notably we have found that the maxima distribution as well as the mean maxima significantly depend on the history, in particular on the value of the previous maximum.

Nevertheless, further work is needed to test if our findings are similar in other (non-Gaussian) initial distributions. In addition, we suggest that memory via the conditional mean maxima and conditional maxima distributions should be considered for improved risk estimation in long-term correlated data. It has been shown recently that multiscaling, which occurs, e. g., in many hydrological time series, has an even more significant impact on risk estimation and the prediction of extreme events like floods (Bogachev, Eichner, and Bunde 2007). Further work is definitely required to study the effects of multiscaling in time series upon extreme value statistics.

Acknowledgement

This work has been supported by the Bundesministerium für Bildung und Forschung and the Israel Science Foundation.

References

Altmann, E. G. and Kantz, H. (2005) Recurrence time analysis, long-term correlations, and extreme events. Phys. Rev. E 71, 056106.

Antal, T., Droz, M., Györgyi, G., and Racz, Z. (2001) 1/f Noise and Extreme Value Statistics. Phys. Rev. Lett. 87, 240601.

Arneodo, A., Bacry, E., Graves, P. V., and Muzy, J. F. (1995) Characterizing long-range correlations in DNA sequences from wavelet analysis. Phys. Rev. Lett. 74, 3293-3296.

Berman, S. M. (1964) Limit theorems for the maximum term in stationary sequences. Ann. Math. Statist. 35, 502-516.

Bogachev, M., Eichner, J. W., and Bunde, A. (2007) The statistics of return intervals in multifractal data sets, preprint.

Bramwell, S. T., Christensen, K., Fortin, J.-Y., Holdsworth, P. C. W., Jensen, H. J., Lise, S., Lopez, J. M., Nicodemi, M., Pinton, J.-F., and Sellitto, M. (2000) Universal fluctuations in correlated systems. Phys. Rev. Lett. 84, 3744-4747.

Bramwell, S. T., Holdsworth, P. C. W., and Pinton, J.-F. (1998) Universality of rare fluctuations in turbulence and critical phenomena. Nature 396, 552-554.

Bunde, A. and Havlin, S. (Eds.) (1994) *Fractals in Science*. Springer, Berlin.

Bunde, A., Havlin, S., Kantelhardt, J. W., Penzel, T., Peter, J.-H., and Voigt, K. (2000) Correlated and Uncorrelated Regions in Heart-Rate Fluctuations during Sleep. Phys. Rev. Lett. 85, 3736-3739.

Bunde, A., Kropp, J., and Schellnhuber, H.-J. (Eds.) (2002) *The Science of Disasters – Climate Disruptions, Heart Attacks, and Market Crashes*. Springer, Berlin.

Bunde, A., Eichner, J. F., Kantelhardt, J. W., and Havlin, S. (2003) The effect of long-term correlations on the return periods of rare events. Physica A 330, 1-7.

Bunde, A., Eichner, J. F., Kantelhardt, J. W., and Havlin, S. (2004) Return intervals ofrare events in records with long-term persistence. Physica A 342, 308-314.

Bunde, A., Eichner, J. F., Kantelhardt, J. W., and Havlin, S. (2005) Long-Term Memory: A Natural Mechanism for the Clustering of Extreme Events and Anomalous Residual Times in Climate Records. Phys. Rev. Lett. 94, 048701.

Chow, V. te (1964) *Handbook of applied hydrology*. McGraw-Hill Book Company, New York.

Corral, A. (2004) Long-Term Clustering, Scaling, and Universality in the Temporal Occurrence of Earthquakes. Phys. Rev. Lett. 92, 108501.

Dahlstedt, K. and Jensen, H. J. (2001) Universal fluctuations and extreme-value statistics. J. Phys. A: Math. Gen. 34, 11193-11200.

Dean, D. S. and Majumdar, S. N. (2001) Extreme-value statistics of hierarchically correlated variables deviation from Gumbel statistics and anomalous persistence. Phys. Rev. E 64, 046121.

Eichner, J. F., Koscielny-Bunde, E., Bunde, A., Havlin, S., and Schellnhuber, H.-J. (2003) Power-law persistence and trends in the atmosphere: A detailed study of long temperature records. Phys. Rev. E 68, 046133.

Eichner, J. F., Kantelhardt, J. W., Bunde, A., and Havlin, S. (2006) Extreme value statistics in records with long-term persistence. Phys. Rev. E 73, 016130.

Eichner, J. F., Kantelhardt, J. W., Bunde, A., and Havlin, S. (2007) On the statistics of return intervals in long-term correlated records. Phys. Rev. E. 75 (in print).

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997) *Modelling Extremal Events*. Springer, Berlin.

Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Camb. Phil. Soc. 24, 180-190.

Galambos, J. (1978) *The Asymptotic Theory of Extreme Order Statistics*. John Wiley and Sons, New York.

Galambos, J., Lechner, J., and Simin, E. (Eds.) (1994) *Extreme Value Theory and Applications*. Kluwer, Dordrecht.

Grissino-Mayer, H. D. (1996) A 2129-year annual reconstruction of precipitation for northwestern New Mexico, USA. In: J. S. Dean, D. M. Meko, and T. W. Swetnam (Eds.), *Tree Rings, Environment, and Humanity. Radiocarbon 1996*, Department of Geosciences, The University of Arizona, Tucson, pp. 191-204.

Guclu, H. and Korniss, G. (2004) Extreme fluctuations in small-world networks with relaxational dynamics. Phys. Rev. E 69, 65104(R).

Gumbel, E. J. (1958) *Statistics of Extremes*. Columbia University Press, New York.

Hurst, H. E., Black, R. P., and Simaika, Y. M. (1965) *Long-term storage: An experimental study*. Constable & Co. Ltd., London.

Kac, M. (1947) On the notion of Recurrence in Discrete Stochastic Processes. Bull. Amer. Math. Soc. 53, 1002-1010.

Kantelhardt, J. W., Koscielny-Bunde, E., Rego, H. H. A., Havlin, S., and Bunde, A. (2001) Detecting long-range correlations with detrended fluctuation analysis. Physica A 295, 441-454.

Kantelhardt, J. W., Penzel, T., Rostig, S., Becker, H. F., Havlin, S., and Bunde, A. (2003) Breathing during REM and non-REM sleep: correlated versus uncorrelated behavior. Physica A 319, 447-457.

Kantelhardt, J. W., Koscielny-Bunde, E., Rybski, D., Braun, P., Bunde, A., and Havlin, S. (2006) Long-term persistence and multifractality of precipitation and river runoff records. J. Geophys. Res. (Atmosph.) 111, 1106.

Koscielny-Bunde, E., Bunde, A., Havlin, S., and Goldreich, Y. (1996) Analysis of Daily Temperature Fluctuations. Physica A 231, 393-396.

Koscielny-Bunde, E., Bunde, A., Havlin, S., Roman, H. E., Goldreich, Y., and Schellnhuber, H.-J. (1998) Indication of a Universal Persistence Law Governing Atmospheric Variability. Phys. Rev. Lett. 81, 729-732.

Koscielny-Bunde, E., Kantelhardt, J. W., Braun, P., Bunde, A., and Havlin, S. (2006) Long-term persistence and multifractality of river runoff records. J. Hydrol. 322, 120-137.

Leadbetter, M. R., Lindgren, G., and Rootzen, H. (1983) *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. (1994) On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Transactions on networking 2, 1-15.

Liu, Y. H., Cizeau, P., Meyer, M., Peng, C.-K., and Stanley, H. E. (1997) Correlations in economic time series. Physica A 245, 437-440.

Majumdar, S. N. and Comtet, A. (2004) Exact Maximal Height Distribution of Fluctuating Interfaces. Phys. Rev. Lett. 92, 225501.

Majumdar, S. N. and Comtet, A. (2005) Airy Distribution Function: From the Area Under a Brownian Excursion to the Maximal Height of Fluctuating Interfaces. J. Stat. Phys. 119, 777-826.

Makse, H. A., Havlin, S., Schwartz, M., and Stanley, H. E. (1996) Method for Generating Long-Range Correlations for Large Systems Phys. Rev. E 53, 5445-5449.

Mandelbrot, B. B. and Wallis, J. R. (1969) Some long-run properties of geophysical records. Wat. Resour. Res. 5, 321-340.

Mann, M. E., Bradley, R. S., and Hughes, M. K. (1999) Northern Hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. Geophys. Res. Lett. 26, 759-762.

Meko, D. M., Therrell, M. D., Baisan, C. H., and Hughes, M. K. (2001) Sacramento river flow reconstructed to A.D. 869 from tree rings. J. Amer. Wat. Resour. Assoc. 37 (4), 1029-1039.

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlén, W. (2005) Highly variable Northern Hemisphere temperatures reconstructed from low- and highresolution proxy data. Nature 433, 613-617. Data from: http://www.ndcd.noaa.gov/paleo/recons.html

Moore, J. J., Hughen, K. A., Miller, G. H., and Overpeck, J. T. (2001) Little Ice Age recorded in summer temperature reconstruction from vared sediments of Donard Lake, Baffin Island, Canada. Journal of Paleolimnology 25, 503-517.

Oliveira J. G. and Barabasi, A.-L. (2005) Darwin and Einstein correspondence patterns. Nature 437, 1251-1253.

Paxson, V. and Floyd, S. (1995) Wide-area traffic: The failure of Poisson modeling. IEEE/ACM Transactions on networking 3, 226-244.

Pelletier, J. D. and Turcotte, D. L. (1997) Long-range persistence in climatological and hydrological time series. J. Hydrology 203, 198-208.

Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. E. (1992) Long-range correlations in nucleotide sequences. Nature 356, 168-170.

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994) Mosaic organization of DNA nucleotides. Phys. Rev. E 49, 1685-1689.

Prasad, R. R., Meneveau, C., and Sreenivasan, K. R. (1988) Multifractal nature of the dissipation field of passive scalars in fully turbulent flow. Phys. Rev. Lett. 61, 74-77.

Rasmussen, P. F. and Gautam, N. (2003) Alternative PWM-estimators of the Gumbel distribution. J. Hydrol. 280, 265-271.

Raudkivi, A. J. (1979) *Hydrology*. Pergamon Press, Oxford.

Raychaudhuri, S., Cranston, M., Przybyla, C., and Shapir, Y. (2001) Maximal Height Scaling of Kinetically Growing Surfaces. Phys. Rev. Lett. 87, 136101.

Rodriguez-Iturbe, I., and Rinaldo, A. (1997) *Fractal River Basins - Change and Self-Organization.* Cambridge University Press.

Schreiber, T. and Schmitz, A. (1996) Improved surrogate data for nonlinearity tests. Phys. Rev. Lett. 77, 635-638.

Shlesinger, M. F., West, B. J., and Klafter, J. (1987) Levy dynamics of enhanced diffusion: application to turbulence. Phys. Rev. Lett. 58, 1100-1103.

Sornette, D. and Knopoff, L. (1997) The paradox of the expected time until the next earthquake. Bull. Seism. Soc. Am. 87, 789-798.

Storch, H. v. and Zwiers, F. W. (2001) *Statistical Analysisin Climate Research*. Cambridge Univ. Press.

Talkner, P. and Weber, R. O. (2000) Power spectrum and detrended fluctuation analysis: application to daily temperatures. Phys. Rev. E 62, 150-160.

Tsonis, A. A., Roebber, P. J., and Elsner, J.B. (1999) Long-range correlations in the extratropical atmospheric circulation: Origins and Implications, J . Climate. 12, 1534–1541.

Turcotte, D. L. (1992) *Fractals and Chaos in Geology and Geophysics*. Cambridge University Press.

Vazquez, A., Oliveira, J. G., Dezsö, Z., Goh, K.-I., Kondor, I., and Barabasi, A.-L. (2006) Modeling bursts and heavy tails in human dynamics. Phys. Rev. E 73, 036127.

Whitcher, B., Byers, S. D., Guttorp, P., and Percival D. B. (2002) Testing for Homogeneity of Variance in Time Series: Long Memory, Wavelets and the Nile River. Wat. Resour. Res. 38 (5), 1054.

Yamasaki, K., Muchnik, L., Havlin, S., Bunde, A., and Stanley, H. E. (2005) Scaling and memory in volatility return intervals in financial markets. Proc. Natl. Acad. Sci. USA 102 (26), 9424-9428.

20 Statistical Properties of Mid-latitude Atmospheric Variability

S. Calmanti[1] , A. Dell'Aquila[2], V. Lucarini[3], P.M. Ruti[4*], and A. Speranza[5]

[1] ENEA, Environmental Department, sandro.calmanti@casaccia.enea.it
[2] ENEA, Environmental Department, alessandro.dellaquila@casaccia.enea.it
[3] Un of Camerino, PASEF, , valerio.lucarini@unicam.it
[4] ENEA, Environmental Department, paolo.ruti@casaccia.enea.it
[5] Un of Camerino, PASEF, Antonio.speranza@unicam.it
[*] Corresponding author

Abstract. This chapter wishes to propose a description of the statistical properties of mid-latitude atmospheric waves for observational data-sets and state-of-the-arts climate models simulations. Therefore, a non-linear theory, accounting for a relevant part of the mid-latitude variability, has been revised in terms of its signature in observational and model data.

Theoretical and observational arguments suggest that two main features of mid-latitude northern hemispheric winter variability can be somewhat unambiguously separated, both in terms of signal and physical processes. The synoptic phenomena, which are travelling waves characterized by time scales of the order of 2-7 days and by spatial scales of the order of few thousands Km., can be associated with release of available energy driven by conventional baroclinic conversion. At lower frequencies (10-40 days), the planetary scale variability is related to the non linear orographic resonance processes

1 Introduction

Non-linear wave self interaction theories (Benzi et al., 1986), predict the existence of multiple equilibria of the mid-latitude planetary wave amplitude. The theory includes the possibility that relatively small changes in the strength of the mean zonal wind (jet strength) may cause a switch from unimodal to multimodal regimes of the atmospheric circulation. Such theories are consistent with the observation that a normally distributed index of the jet strength (Malguzzi and Speranza, 1981) correspond to a non-normal distribution of any index related to the planetary wave activity (Hansen and Sutera, 1986). The statistical significance of the null-hypothesis that distribution of wave activity indexes is non-normal is still debated (Christiansen, 2005; Stephenson et al. 2004). The discussion includes the possibility that considering the distribution of atmospheric activity indexes as the results of the clustering around fixed point may be just insufficient to describe the complexity of the atmosphere so that other approaches should be attempted. However, the notion that well-defined winter mid-latitude atmospheric patterns of flow are recurrent during the northern hemispheric winters (Dole, 1983) has been repeatedly put forward, investigated and

debated since the early definition of Grosswetterlage (Baur, 1951), to the classical identification of Atlantic blocking (Rex, 1950), all the way to more recent work on regimes detection and identification (Corti et al., 1999).

The atmospheric wave statistical properties evidence, not only a low frequency component (period of 10-40 days), but a baroclinic component with eastward propagating waves and well defined phase-relationship. Since the milestone paper of Charney and De Vore (1979), the low frequency component of the mid-latitude atmospheric variability has been related to orographic resonance processes, and discussed in terms of mid-latitude atmospheric regimes. The synoptic travelling waves characterized by time scales of the order of 2-7 days and by spatial scales of the order of few thousands Km., are associated with release of available energy driven by conventional baroclinic conversion (Blackmon, 1976; Speranza, 1983; Wallace et al, 1988). The dynamical interaction between these two distinct spectral regions has been deeply investigated and debated during 80's (see K. Haines, 1994, for a review), but a complete understanding of their mutual nature is still lacking.

Building upon this conceptual framework, the observational evidence of the bent resonance curve obtained in non-linear wave self interaction theories by Benzi et al. (1986) has been revised, and the relationship with the strength of the westerlies has been focused on (section 2, "Non linear behaviour of the planetary waves"), taking advantage of a longer observational basis, consisting of the two major reanalysis products released by NCEP-NCAR (Kistler, et al, 2001) and by ECMWF (Simmons and Gibson, 2000).

Upon the evidence of the non-linear behaviour of the planetary waves, we want to face the capability of state of the art climate models for reproducing this relevant feature.

In fact, the operational global weather forecasting models, which constitute the bases of the atmospheric components presently adopted in up-to-date global climate models, routinely used in the '80s were well-known to suffer from serious biases in the representation of the northern hemisphere winter mid-latitude synoptic and planetary atmospheric disturbances (Tibaldi, 1986). In particular, it was confirmed in a number of studies that such models on the average featured a large overestimation of the baroclinic synoptic waves (e.g. Sumi and Kanamitsu, 1984; Klinker and Capaldo 1986; Siegmund 1995) and a large underestimation of the planetary waves (e.g. Wallace et al., 1983), with biases sometimes of the same size as the average climatology of the observations and analyzed atmospheric fields. Some brief notes on the history of atmospheric modelling can be found at http://www.aip.org/history/sloan/gcm/intro.html.

A relevant question should be answered: do state of the art climate models suffer from similar biases, and if this is the case, how these biases impact on the representation of the non-linear behaviour of the planetary waves?

A spectral analysis of the mid-latitude waves propagation reveals that climate models can properly describe a linear atmosphere, i.e. the atmospheric waves have the correct frequency/wave-number and amplitude. In order to analyze this problem, we first characterize the statistical properties of the mid-latitude variability in the spectral space explored by means of the Hayashi decomposition (Hayashi, 1971, 1979; Pratt, 1976; Fraedrich and Bottger, 1978). We use the 500hPa geopotential

height which is one of the most relevant variables descriptive of the large scale atmospheric circulation (Blackmon, 1976), both for the reanalysis products (section 3, "Mid-latitude Northern Hemispheric waves variability") and the global climate models, contributing to the scientific basis of the IPCC 4th Assessment Report (section 4, "Statistical properties of the mid-latitude waves in the IPCC models"). However a realistic description of the spectral properties does not imply a realistic description of the wave-wave interaction mechanisms. A further wave-analysis has been performed in order to verify the consistency of the climate models respect to the wave-self interaction theory.

2 Non-linear Behaviour of the Planetary Waves

Since the pioneering work of Charney and Devore (1979), the low frequency component of the mid-latitude atmospheric variability has been related to orographic resonance processes. Charney and DeVore (1979) studied the dynamics of planetary waves wind-driven excitation due to orography and mathematically closed the system by introducing a form-drag relationship relating the zonal wind evolution to the amplitude of the waves themselves. A minimal truncation of the motion equations featured the existence of two different stable equilibria for the system characterized by two distinct pair of values for the wave amplitude and for the zonal flow. The observational evidence of a unimodal zonal wind imposed some modifications of the Charney and DeVore theory. Benzi et al. (1986) and Speranza (1986) proposed a model in which the nonlinearity is introduced in the form of wave self-interaction. The wave self-interaction causes the modification of the linear wave-wind resonance curve by introducing, along the lines of the classical examples due to Landau in mechanics (Landau and Lifshitz, 1976) and statistical physics (Landau and Lifshitz, 1980), a bent resonance curve, which allows different states of the wave amplitude corresponding to the same value of the jet strength, above a certain threshold. In Ruti et al. (2006), the statistical properties of the planetary waves of the real atmosphere have been demonstrated to have a critical dependence on the strength of the upper tropospheric jet. Here, we expose in more details the mentioned findings.

2.1 Dynamical indexes

The real atmosphere is a complex and complicated dynamical system featuring chaotic dynamics on all spatial and temporal scales. Therefore, in order to define reasonable counterparts to the dynamical parameters used in Benzi et al. (1986), we need to introduce robust indicators providing a picture of the amplitude of the planetary wave and of the strength of the jet.

Following Hansen and Sutera (1986) and Benzi and Speranza (1989), we average the daily 500 hPa geopotential height over the latitudinal belt, 37°N-72°N. We then compute for each day the spatial discrete Fourier transform of the field, and define the so called wave activity index (WAI) as follows:

$$WAI(t) = \left(\sum_{k=2}^{4} 2 \cdot |A_k(t)|^2 \right)^{1/2}, \quad (1)$$

where $|A_k(t)|$ is the Fourier coefficient of the wave having wavenumber k at the time t. As thoroughly discussed in the literature, such an index is aimed at providing a synthetic picture of the planetary wave activity and summarizing the overall atmospheric response at ultra-long wavelengths. The WAI index has been introduced with the aim of capturing the topographic resonances, since the dominant wave-number of the mid-latitude northern hemisphere topography ranges from 2 to 4. The Hayashi spectra confirm that the bulk of variability is concentrated in the zonal wavenumbers k=2,3,4 (Blackmon, 1976; Speranza, 1983; Dell'Aquila et al., 2005). In order to attenuate the direct influence of the seasonal cycle we erase from the signal its first 4 harmonics of the $^{1}y^{-1}$ frequency. Along the same lines, in order to mitigate the signature of the synoptic atmospheric variability, we apply a low-pass FFT 8-day filter to the signal.

We also introduce the jet strength index (JSI) as the signal constituted by the daily maximum values of the field given by the zonal average of the 200hP zonal wind. We emphasize that such maximum always results to be realized between 30°N and 35°N. In order to mitigate the signature of the synoptic atmospheric variability, we then apply a low-pass filter by performing a 8-day FFT filter to the signal. Since in the context of the theory developed by Benzi et al. (1986) the jet strength is considered as an autonomous forcing parameter of the system, we do not filter out from the JSI the seasonal cycle and its harmonics as done for the WAI. Obviously, in the context of a linear reasoning, the cancellation of these frequencies performed of the WAI hides part of the influence of the jet strength, but we assume that such effect is noticeably smaller than the direct influence of the seasonal cycle itself. Eventually, we extract for both indicators WAI and JSI the winter data by considering for all years the days between December 1st and February 28th.

2.2 Bimodal statistical signature

When considering the overlapping period 1958-2002, the NCEP and ERA 40 datasets compare well in terms of the statistics of the WAI and JSI time-series. The one-dimensional (1D) Kolmogorov-Smirnov (K-S) test gives in both cases a highly statistically significant agreement. The agreement is excellent if we consider the time series of the deviations from the mean value. Similar results are obtained when the two-dimensional (2D) [JSI, WAI] time-series of the NCEP and ERA 40 datasets are compared with the stricter 2D K-S test (Peacock, 1983; Fasano and Franceschini, 1987). In agreement with previous suggestions (Hansen and Sutera 1986, Benzi and Speranza 1989), while both the NCEP and ERA 40 JSI time series are compatible in terms of K-S test with a Gaussian signal with corresponding mean and standard deviation, the WAI time series is markedly non-Gaussian. The 2D [JSI, WAI] time-series of both reanalyses result to be not compatible with bivariate Gaussian signals.

We then construct, with the help of two-dimensional Gaussian kernel estimators the empirical pdf of the 2D [JSI, WAI] time-series. Figure 20.1 show the joint 2D pdfs for the NCEP and ERA40 re-analysis data. We notice quite clearly that, for low values of JSI (38<JSI<41m/s), the atmosphere populates a state characterized by low or intermediate values of WAI, while for higher values of JSI (42<JSI<47 m/s), the atmospheric phase-space shows a more complex structure, with the appearance of a minimum between low and high values of the WAI. Considering higher values of the JSI, an atmosphere characterized by low values of WAI appears. In the ERA40 this feature is less evident, but nevertheless present. The JSI acts as parameter of a dynamical system setting the complexity of the pdf of the planetary waves indicator.

The nonlinear picture outlined above poses some questions relatively to oceanic forcing. In particular, which is the position of the El Nino and La Nina years in our 2-D phase-space?. The main El Nino years (57/58, 72/73, 82/83, 97/98) produce a stronger and elongated Pacific jet respect to normal conditions (the opposite in the La Nina case), and the corresponding JSI is more than 47 m/s locating the El Nino years in the upper region of the phase-space where the nonlinear behavior is less evident. Two La Nina events (75/76 and 88/89) have low values of the JSI, populating the lower part of the 2-D phase space. The other two strong La Nina events (70/71 and 73/74) and the El Nino 65/66 event attain a value of JSI in the range 42-47 m/s.

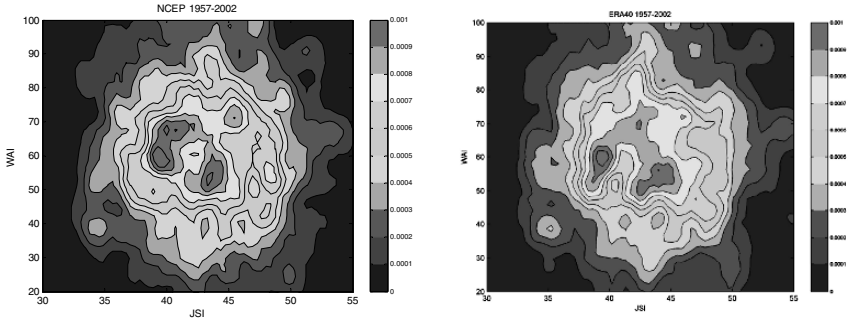


Fig. 20.1. Evidence of the dependence of the planetary wave regimes on the jet strength for the winter-time Northern Hemisphere. Illustrated is the two-dimensional joint pdf using the index of the planetary waves (WAI) and the index of the tropospheric jet strength (JSI). The period considered is the 1958-2002 time frame for the NCEP and ERA40 datasets.

By clustering our 2D space for low and high JSI values, we can obtain a pair of 1D kernel pdfs, each comprising, for both reanalyses, about 1/3 of the total population (approximately 1400 over 4050 days). Figure 20.2 shows the empirical pdfs obtained with a Gaussian kernel estimator (Silverman, 1986) with the NCEP data for the subsets of the WAI time series obtained by selecting the days respectively characterized by the above defined low and high JSI values. For low and high JSI values, we observe a unimodal distribution, while, for intermediate JSI values, a bimodal distribution appears.

The relevance of the characterization of the atmospheric phase-space for climate change studies (Corti et al., 1999) requires statistical robustness in the properties of the pdfs in order to avoid artificial results (Stephenson et al., 2004). Based on the statistical analysis in Ruti et al. (2006) the results can be evaluated safe and robust.

3 Mid-latitude Northern Hemispheric Waves Variability

In the previous paragraph, we considered the statistical distribution of a planetary index characterizing the activity within the low-frequency range (10-40 days). Here, we extend the analysis to the complete frequency range, considering scales from synoptic to planetary. In fact, such scales characterize the statistical properties of the mid-latitude circulation, and, in turn, the climate of the system (in this case, the Northern Hemispheric mid-latitudes). It is, at this point, of paramount importance settling out the "best" statistical properties of the system, we have. So, reanalysis products (i.e., NCEP-NCAR, Kistler et al, 2001, and ECMWF, Simmons and Gibson, 2000; hereafter referred as NCEP and ERA40) which are designed to obtain global, homogeneous and self-consistent datasets of the atmospheric dynamics on the

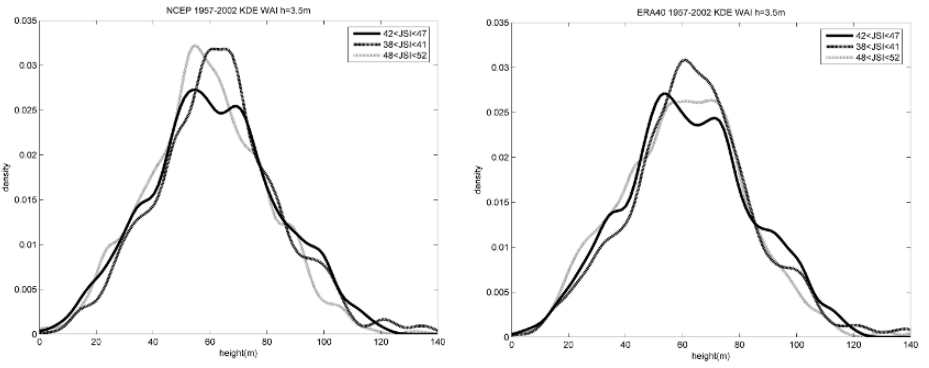


Fig. 20.2. Results from clustering our atmospheric phase-space in low, intermediate, and high values of the tropospheric jet strength (JSI). Illustrated is the one-dimensional pdf of the planetary waves index (WAI) for low JSI (red), intermediate (blue), and for high JSI (black). The period considered is the 1958-2002 time frame for the NCEP and ERA40 datasets.

longest time scale allowed by the currently available instrumental data (Kistler et al., 2001), become our "best" guess of the climate system.

The re-analysed data have been used in a large variety of contexts, ranging from the direct cross validation with independent observations (Josey, 2001; Renfew et al., 2002), to the study of specific aspects of atmospheric dynamics (Annamalai et al., 1999; Hodges et al., 2003; Dell'Aquila et al, 2005; Dell'Aquila et al., 2006; Ruti et al., 2006), to their usage as external input for the modeling of other components of the climate system (Harrison et al., 2002; Josey et al., 2002). Since they represent the best available homogeneous datasets for climate studies and for validation of climate general circulation models (GCMs), it is quite relevant to assess the consistency of these reanalyses in the representation of the variability of the atmosphere.

3.1 The climatology of the Hayashi decomposition

The 500hPa geopotential height is a very powerful proxy for characterizing atmospheric variability on a wide range of spatial and temporal scales (Blackmon, 1976). So, we use the freely available northern hemisphere 500hPa geopotential height reanalysis provided by the NCEP and ERA 40 datasets. We consider the datasets for the overlapping time frame ranging from September 1st 1957 to August 31st 2002. Both reanalyses are publicly released with spatial resolution of 2.5° x 2.5°, with a resulting effective horizontal grid of 144 x 73 points. The maximum time resolution is six hours. However, in order to focus on time scales longer than one day, for both reanalyses we consider daily data, obtained by arithmetically averaging the available 4 times daily data.

Our study focuses on the northern hemisphere mid-latitude atmospheric winter variability. To this aim, we follow Benzi and Speranza (1989) and select the December-January-February (DJF) data relative to the latitudinal belt 30°N-75°N where the bulk of the baroclinic and of the low frequency planetary waves activity is observed. We average day wise the geopotential field over such latitudinal belt in order to derive a one dimensional longitudinal field representative of the atmospheric variability in the mid-latitudes.

The variability of the one dimensional field in terms of components of different periods and wavelengths can be effectively described by means of the space-time Fourier decomposition introduced by Hayashi (1971, 1979). By computing the cross-spectra and the coherence of the signal, such method allows for a separation of the eastward and westward wave propagating components from the standing component. In the Dell'Aquila et al. (2005), we present a full account of this methodology.

Figures 20.3a-3d show the various components of the 45-winters averages of the Hayashi spectra as computed from the NCEP reanalysis dataset. The spectra express the energy density of the wave field with respect to frequency and wavenumber and its decomposition into standing and propagating components. Figure 20.3a shows $\overline{H}_T(k,\omega)$, the total energy spectrum, Fig. 20.3b shows $\overline{H}_S(k,\omega)$, the energy spectrum related to standing components, Fig. 20.3c shows $\overline{H}_E(k,\omega)$, the energy spectrum related to eastward propagating components, and Fig. 20.3d shows $\overline{H}_W(k,\omega)$, the energy spectrum of the westward propagating components. The overbar indicates the operation of averaging over the N=45 winters. Throughout the chapter, the indexes T, S, E, and W indicate total, standing and eastward/westward propagating components, respectively. As customary in literature, the Hayashi spectra have been obtained by multiplying the energy spectra by $k \cdot \omega / 2\pi$, in order to compensate for the non-constant density of points in a log-log plot.

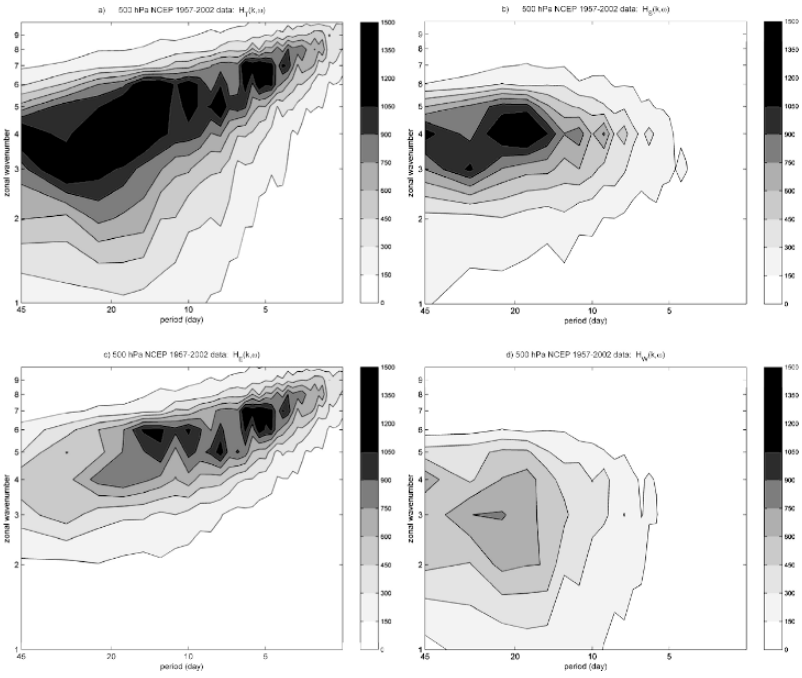


Fig. 20.3. Climatological average over 45 winters of Hayashi spectra for 500 hPa geopotential height (relative to the latitudinal belt 30°N-75°N) from NCEP data: $\overline{H}_T(k,\omega)$ (a); $\overline{H}_S(k,\omega)$ (b); $\overline{H}_E(k,\omega)$ (c); $\overline{H}_E(k,\omega)$ (d). The Hayashi spectra have been obtained multiplying the energy spectra by $k \cdot \omega / 2\pi$. The units are m2. (Adapted from Dell'Aquila et al, 2005).

A large portion of the total energy is concentrated in the low frequency – low wavenumber domain, and can be related mostly to standing waves and to westward propagating waves, as can be deduced from Figs. 20.3b and 20.3d. The high frequency - small wavelength domain, corresponding mainly to synoptic disturbances, contains a smaller portion of the total energy, and is essentially related to eastward propagating waves. The fact that eastward propagating waves are essentially characterized by short wavelength and that long wavelengths characterize westward propagating waves is consistent with the Rossby waves picture. Accordingly, in Fig. 20.3c we observe that it is possible to recognize the frame of a monotonic dispersion relation $\omega = \omega(k)$ for eastward propagating waves. Instead, in Fig. 20.3d the appearance of a dispersion relation for westward propagating waves is unclear. Since the long waves do not seem to obey even statistically a well-defined dispersion relation, it is more correct to refer to these spectral components with the more generic term of disturbances.These results have a good agreement with the past analyses performed along the same lines (Fraedrich and Bottger, 1978; Speranza, 1983).

A first insight of the discrepancies between the two reanalyses can be gained by considering the difference between the two mean Hayashi spectra, portrayed in Figs. 20.4a-20.4d.

When considering the difference of the two total energy spectra (Figure 20.4a), major discrepancies can be observed in the high-frequency range, corresponding to synoptic, eastward propagating disturbances. This feature is confirmed in Fig. 20.4c where a significant discrepancy is observed in this component of the Hayashi decomposition. Instead, no significant difference is found in the westward propagating signal (Figure 20.4d).

When considering the difference spectrum of the standing component energy (Figure 20.4b), we observe that the most significant contribution comes from the low frequency-low wavenumber sector. In particular, large differences are observed for wavenumbers ranging from 3 to 5, and for periods longer than 20 days. In this domain we do not observe large discrepancies for the total energy spectra. This implies that the two datasets should differ in the repartition of the energy between standing and traveling components.

Such disagreements are quite impressive because the Hayashi spectra have been averaged over the 45 winters, thus proving that the climatology of the winter atmospheric variability of the two datasets is not fully equivalent.

3.2 The interannual variability.

In this paragraph we inspect the temporal behavior of the hayashi decomposition by considering specific spectral subdomains. We introduce the following integral quantities:

(2) $E_j^n(\Omega) = \int_{\omega_1}^{\omega_2}\int_{k_1}^{k_2} d\omega dk \; H_j^n(k,\omega)$, with j=T,S,E,W;

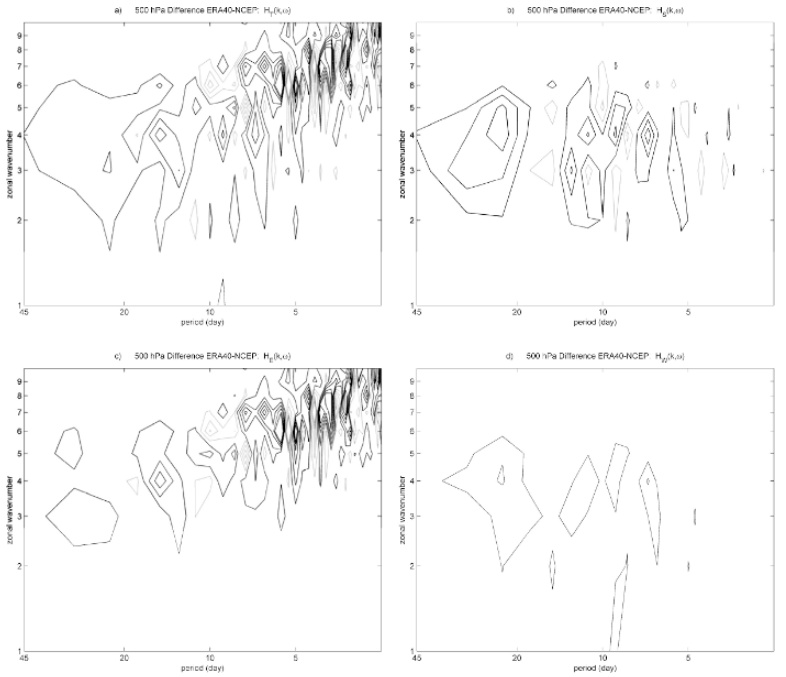


Fig. 20.4. As in Fig. 20.3 but for the difference between ERA40 and NCEP dataset Hayashi spectra. Black contours pertain to positive values and grey contours to negative values. The units are m2. Contour interval is 20 m2. We do not plot the zero contours. (Adapted from Dell'Aquila et al, 2005).

where n indicates the year. The integration extremes, $\omega_{1/2}$ and $k_{1/2}$, determine the spectral region of interest $\Omega = [\omega_1, \omega_2] \times [k_1, k_2]$. The quantity $E_j^n(\Omega)$ introduced in equation (1) represents the fraction of energy of the full spectrum associated to a given subdomain Ω and to a given year n. Therefore, for various choices of Ω, the com parison of the quantities $E_j^n(\Omega)$ obtained from the two reanalyses can help identifying differences in the capability of describing processes occurring on a given spatial and temporal scale and pertaining to qualitatively different physical processes.

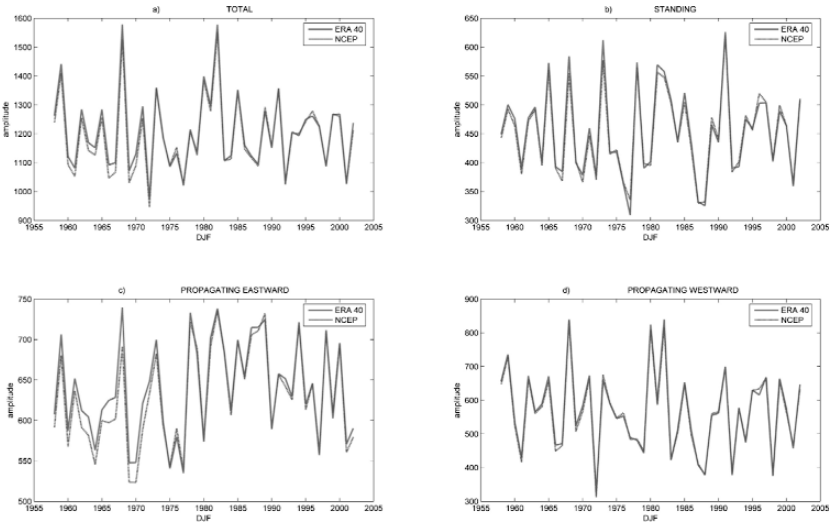


Fig. 20.5. Time series of the quantity $E_j^n(\Omega)$ computed for the two reanalysis datasets, for categories Total , Standing, Eastward and Westward spectral components.

Inspection of the inter-annual variability in $E_j^n(\Omega)$ for the total, standing and eastward/westward propagating components (Figure 20.5), reveal no evident trends, but for a change in the eastward component since 1979. The later seems to suggest an increase in the second part of the record of the variance associated with the eastward waves. The introduction of satellite data since 1979 (and partially since 1973) could be related to this change, in fact discrepancies between the two re-analysis in the baroclinic spectral range have been found (Dell'Aquila et al., 2005).

4 Statistical Properties of the Mid-latitude Waves in the IPCC Models

In view of the remarks in section 2 and 3, we compare the performances of 17 global climate models, contributing to the scientific basis of the IPCC 4th Assessment Report (henceforth, IPCC-4AR), in the description of the northern hemisphere mid-latitudes winter atmospheric disturbances by analysing in detail the statistical properties of the modelled 500hPa geopotential height fields (Lucarini et al., 2006). In addition to the issue of self-consistency, we assess the realism of the climate models by comparing their statistics with those of the NCEP-NCAR and ECMWF global reanalyses, which have been analyzed in detail elsewhere (Sec. 2 and Dell'Aquila et al., 2005). We perform for all models a space-time spectral analysis devised along the way of Hayashi (1971; 1979), which allows the definition of the spectral densities

of the all sort of atmospheric disturbances as well as the separation of the spectral densities describing statistically standing and propagating waves. See May (1999) for an example of application of Hayashi spectra techniques for the assessment of the performances of climate models. To support the tantalizing aim of planning practical strategies for the improvement of numerical climate models, the Project for Climate Model Diagnostics and Intercomparison (PCMDI: http://www-pcmdi.llnl.gov) has gathered the output of climate models produced worldwide into a single location and solicited the provision of simple scalar metrics of model performances. We underline the possibility of distinguishing two qualitatively distinct, albeit related, kinds of metrics. The comparison of bulk thermodynamic quantities defining the climate state, such as the tropospheric average temperature, tropospheric average specific humidity, variance of geopotential height, allows the definition of global metrics which may be considered as robust diagnostic tools. Nevertheless, such an approach does not allow for the disentanglement of the role of each one of the vast range of distinct physical processes contributing to the global balances. In order to capture the differences in the representation of specific physical processes, it is necessary to use specialized diagnostic tools - that we may call process-oriented metrics - as indexes for model reliability. Such approach may be helpful in clarifying the distinction between the performance of the models in reproducing diagnostic and prognostic variables of the atmospheric system.

In the present analysis, we introduce the global and process oriented metrics by defining ad hoc integral measures over the different spectral sub-domains defined in the previous section. The estimate of the total wave variability of the field is chosen as a global scalar metrics assessing the overall statistical properties of all sort of atmospheric disturbances. The estimate of the variability pertaining to the eastward propagating HFHW (period ranging since 2 to 7 days, wave numbers since 6 to 72) waves and to the standing LFLW (period ranging since 10 to 45 days, wave numbers since 2 to 4) waves (see Table 1) are taken as scalar metrics, accounting phenomenologically for the model representation of specific physical processes and specifically related to the capability in describing the travelling baroclinic and planetary disturbances, respectively (see section 3). We maintain that, by applying this method, the reliability of the considered models in the simulation of some basic climatic processes can be critically addressed.

Table.20.1. Definition of 4 regions in the Hayashi spectra of the winter atmospheric variability. LFLW: Low Frequency Long Wavenumber; LFHW: Low Frequency High Wavenumber; HFLW: High Frequency Long Wavenumber; HFHW: High Frequency High Wavenumber. The values $\omega_2 = 2\pi/2d$, $k_2 = 72$ constitute the highest frequency and wavenumber allowed by the adopted data resolution.

Spectral properties	$k_1 = 2, k_2 = 4$	$k_1 = 6, k_2 = 72$
$\omega_1 = 2\pi/45d$, $\omega_2 = 2\pi/10d$	$\Omega = \mathrm{LFLW}$	$\Omega = \mathrm{LFHW}$
$\omega_1 = 2\pi/7d$, $\omega_2 = 2\pi/2d$	$\Omega = \mathrm{HFLW}$	$\Omega = \mathrm{HFHW}$

The IPCC models considered in this study are described in Lucarini et al. (2006) along with their main features. The simulated time frame considered for the comparison is 1962-2000. For this period a control simulation has been performed by all IPCC models, imposing the observed concentration of trace gases, such as CO2 and ozone.

The IPCC models output is compared with observations by considering two major reanalysis datasets: NCEP and ERA40.

Since the goal of this study is to diagnose the mid-latitude winter atmospheric variability of the considered climate models, along the lines of Dell'Aquila et al. (2005), we should use the December-January-February (DJF) daily values of geopotential height at 500hPa data averaged over the latitudinal belt 30°N-75°N.

Unfortunately, this field is not one of the standard daily 3D field outputs for the IPCC models, which comprises only zonal and meridional wind speed, air temperature and specific humidity.

We have, so, retrieved the 500 hPa meridional wind speed, and making use of the geostrophic relationship (see Lucarini et al., 2006) we compute the 500 hPa geostrophic geopotential height.

4.1 Global metrics

The space-time spectra is computed for each of the 39 winters included in our datasets, using the hayashi decomposition technique (see section 3).

A global scalar metrics can be introduced by integrating the whole space-time power spectrum corresponding to each one of the 39 considered winter seasons. We then consider the quantity $E_T^n(\Omega)$ introduced in equation (2) where Ω is set to be the full frequency-wavenumber domain and n is the index running over the considered winters.

In Fig. 20.6 we show for all the considered datasets the climatological average $\overline{E}_T(\Omega)$of the integral of the full spectrum versus its interannual variability $\sigma_{E_T(\Omega)}$ (Standard deviation). In this scatter diagram the abscissas represent the model average intra- seasonal variability, while the ordinates represent the interannual variability of the model intra-seasonal variability. The models ensemble average is also shown. Such scatter diagram allows a complete visual representation of a given climate by reporting the two most relevant statistical moments of any chosen variable.

As shown in section 3, the reconstructed latitudinally averaged 500 hPa geopotential height fields of the two reanalyses have similar intra-seasonal variability and slightly different interannual variability.

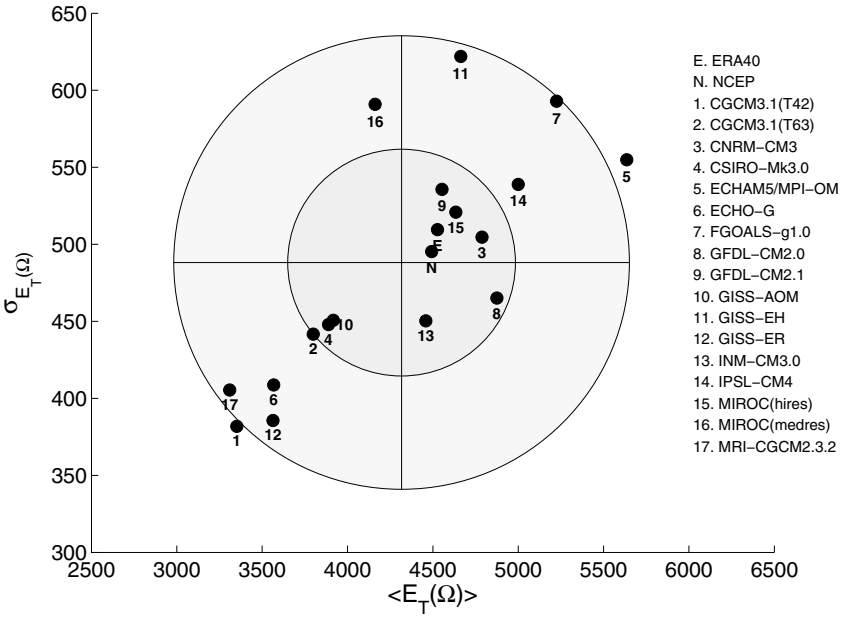


Fig. 20.6. Mean annual full spectrum integral $\overline{E}_T(\Omega)$ and its variance for the 2 reanalyses and the 17 GCMs. The letters indicate the data computed from the NCEP-NCAR reanalysis (N), from the ERA40 reanalysis (E). The shaded areas represent the dispersion of data: the center of the ellipses is the ensemble average; the semi-axes of the inner ellipse are equal to the variance of data in the corresponding direction; the semi-axes of the outer ellipse correspond to twice the variance. (Adapted from Lucarini et al., 2006).

Apparently, if we consider the model ensemble, we have a good agreement with observations. Both reanalyses lie well within one standard deviation from the ensemble average (center of the ellipses in Fig. 20.6). However, the models are widely spread over the plane space spanned by the two variables, with a typical bias of about 15% with respect to the reanalyses. In particular, considering the biases on the quantity $\overline{E}_T(\Omega)$, and considering that the standard deviation of the climatological mean $\overline{E}_T(\Omega)$ can be approximated as $\sigma_{E_T(\Omega)}/\sqrt{N} \approx 0.16\sigma_{E_T(\Omega)}$, since the winters are weakly dependent, we have that very few models are statistically consistent with the reanalyses with a reasonable significance.

In this sense, the best models are by far the high-resolution version of the MIROC and the GFDL-CM2.1. The T62 version of the CGCM3.2 and the relatively low-resolution INM3.0 also perform well, featuring a slightly too large intraseasonal

variability and too low interannual variability, respectively. Some models, such as FGOALS1.0 and ECHAM5/MPI-OM, feature very large positive biases ranging over 20% for the interannual and intraseasonal variability. Other models, usually of relatively low-resolution, such as CNRM-CM3 GISS-ER, ECHO-G and MRI-CGCM2.3.2, feature over 20% negative biases. In general, the biases on the intraseasonal and interannual variability are positively, linearly correlated: for larger average signals the variability tends to be larger, so that the model spread in Fig. 20.6 is definitely non-isotropic.

In some cases, it is possible to track the improvements occurring between different versions or set-ups of the same climate model. Thusthe GISS-EH model (which includes an isopycnal ocean component) has a better representation of the intraseasonal atmospheric with respect to the GISS-ER which has a z-coordinate ocean model. However, the two models have considerably different interannual variability, which is in both cases different from that of the reanalysis. The CGCM3.1 model is presented in two versions which are identical except for the horizontal resolution (T47, corresponding to about 3.75° resolution, and T63 corresponding to about 2.8° resolution). The CGCM3.1(T63) improves the representation of both the intraseasonal variability and of the interannual variability with respect to CGCM3.1(T47). Similar improvements are observed between MIROC (high resolution) and MIROC (medium resolution) and between GFDL-CM2.1 and GFDL-CM2.0. The MIROC model is presented with two different horizontal as well as vertical resolution (T42L20 for the medium resolution version and T106L56 for the high resolution version). The medium resolution version shows already quite good performances and is among the best models. However, a substantial improvement is observed when switching to higher resolution. The GFDL-CM2.0 and GFDL-CM2.1 models are very similar. In particular they share the same horizontal as well as vertical resolution, but in the GFDL-CM2.1 model some numerical techniques are improved with respect to GFDL-CM2.0. For example, the horizontal in CM2.0 uses centered spatial differencing, whereas in CM2.1 the hori zontal discretization is performed with a flux-form semi-Lagrangian method. Finally, let us note that the models with a better representation of intraseasonal and interannual variability display also the most realistic ENSO variability as shown in Van Oldenburg et al. (2005).

4.2 Process oriented metrics

In order to construct a process-oriented metrics pointing at the diagnostics of specific physical processes, we consider the decomposition of the whole frequency-wave number domain, proposed in Table 1, into two spectral subdomains: the LFLW subdomain, which includes periods from 10 to 45 days and zonal wavenumbers $2 \le j \le 4$ (length scales larger than 7000Km); the HFHW subdomain, which includes periods from 2 to 7 days and zonal wavenumbers $j \ge 6$ (length scales ranging from a few hundreds of kilometres to 5000Km). For each year of a given dataset, we then provide a bulk measure of the planetary standing waves and of the eastward propa-

gating baroclinic waves by $E_S^n(\Omega_{LFLW})$ and $E_E^n(\Omega_{HFHW})$ defined in equation (2), respectively, where the two Ω-domains are prescribed as above, n is the index of the winters, and the lower indexes S and E refer to standing and eastward propagating components, respectively.

The quantities $E_S^n(\Omega_{LFLW})$ and $E_E^n(\Omega_{HFHW})$ can then be used to characterize the abilities of each model incorrectly describing two specific physical processes of the atmosphere.

We focus on the average (climatological) description provided by each model of the planetary standing spectral component and of the baroclinic eastward propagating waves and keep slightly aside the problem of checking the correctness in the description of the interannual variability of the two signals. We then consider for each model the quantities $\overline{E}_S(\Omega_{LFLW})$ and $\overline{E}_E(\Omega_{HFHW})$ with the corresponding standard error estimated as expressed as $\sigma_{E_S(\Omega_{LFLW})}/\sqrt{N}$ and $\sigma_{E_E(\Omega_{HFHW})}/\sqrt{N}$, respectively. This entails assuming a weak time-lagged correlation between the 39 winter seasons for the two signals in each model, which is correct in the first approximation.

In Fig. 20.7 we show for all the considered datasets the climatological average $\overline{E}_S(\Omega_{LFLW})$ versus the climatological average $\overline{E}_E(\Omega_{HFHW})$, and indicate for both directions the corresponding standard error. When considering the ERA40 and NCEP renalayses datasets, we observe that the two reanalyses are very close to each other and they could be brought to a closer agreement by dropping the pre-satellite period (Dell'Aquila et al., 2005). However, the discrepancies among IPCC models by far exceed the discrepancies between the two different dynamical interpolations of the available observations.

In particular, with this process oriented metrics, the agreement with observations seems weaker than with the global metrics, with the two reanalysis lying at just about one standard deviation from the ensemble average. It is then important to stress that with the global metrics, the errors originating in different regions of the spectrum may average out and hide major model deficiencies. Instead, in with the process oriented metrics, it is easier to bring the model's biases into light.

For example, a general (with very few exceptions) occurrence for all IPCC models is that they overestimate the variability in the HFHW subdomain. Regarding the LFLW subdomain, more than half of the models tend to underestimate the corresponding variability. As an overall result, the model ensemble (center of the ellipses in Fig. 20.7) overestimates the HFHW variability and underestimates the LFLW variability. This closely resembles the results of many diagnostic studies performed in the past on global weather forecasting models (Tibaldi, 1986), which constitute the ancestors of the atmospheric components of the considered climate models.

Moreover, if we consider the standard deviation of the climatological means we again have that very few models are statistically consistent with the reanalyses with a

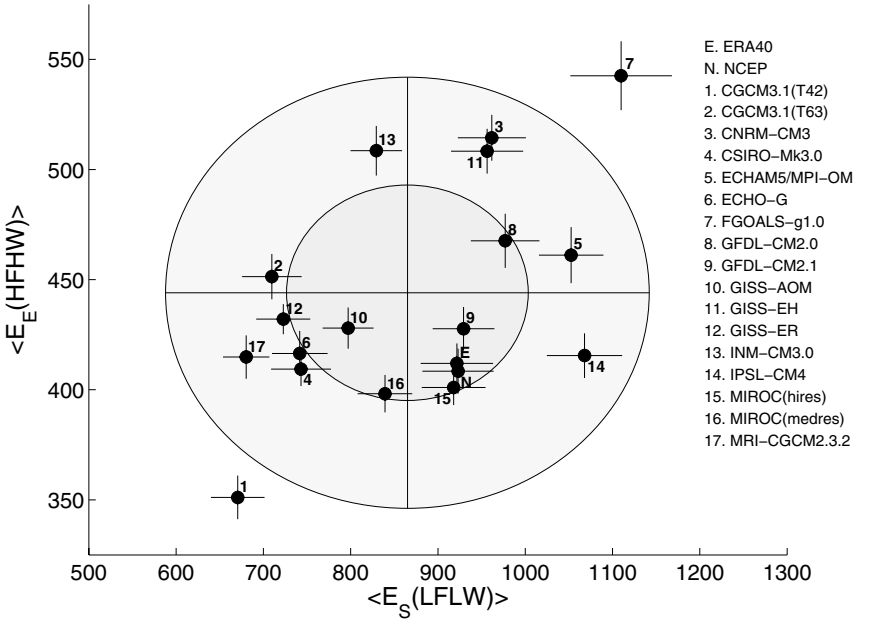


Fig. 20.7. As in Fig. 20.6 but for the low-frequency low-wavenumber subdomain $\overline{E}_S(\Omega_{LFLW})$ (abscissas) of the standing disturbancesversus the high-frequency high-wavenumber subdomain $\overline{E}_E(\Omega_{HFHW})$ (ordinates) of the eastward propagating component. (Adapted from Lucarini et al., 2006).

reasonable significance. The models that are closer to the reanalyses are the MIROC(hires) model and the GFDL2.1 model. In these two cases, the error bars have some overlap with the area defined by the error bars of the NCEP and ERA40 reanalyses. By using this metrics, the models that appear to have the worst performances are the CNRM-CM3 model, which underestimates both the low frequency and the high frequency, and the FGOALS model, which overestimates the variability in both spectral sub-domains.

Unfortunately, there is no unique way to attribute the biases of LFLW and HFHW variability to specific features that are common to a class of models. Instead, there are cases in which different versions of the same model show different behaviour in this spectral subdomain. We refer to the CGCM3.1-T47 and CGCM3.1-T63 models, which differ for the horizontal resolution, and the GISS-EH and GISS-ER models, which share the same atmospheric component over different ocean models. Therefore, it is interesting to compare the relative changes in the performances of models presented with different configuration. This might help in the identification of specific strategies for model improvements.

In the case of the CGCM3.1 models, an increased horizontal resolution (T63) leads to a better statistics of planetary-scale standing waves. However, the performance on small scale disturbances is better in the low resolution version (T47). Thus,

increasing the horizontal resolution alone leads to no automatic overall substantial improvement of the model performance.

As mentioned above, the two versions of the GISS model share the same atmosphere component over different oceans. In this case, the use of vertical density coordinates (GISS-EH) appears to improve the statistics of the standing waves which is too low in the case of z-coordinates (GISS-ER). Instead, the performance in on eastward propagating variance is worse.

In the case of the MIROC model, increased resolution alone improves the model performances. However, unlike with the CGCM3.1 models, in this case both the horizontal and the vertical resolution is increased.

Also in the context of process-oriented metrics, the improvement from GFDL-CM2.0 to GFDL-CM2.1 is substantial, with the latter in good agreement with the reanalyses. The experience with the GFDL models demonstrates how the increase of computer power (i.e. of resolution) may not be the only pathway to model improvements. Instead, the study of more accurate discretization techniques may give a substantial contribution.

4.3 Planetary disturbances in GCMs

The model simulations, gathered by the IPCC, reveal a reasonable skill in reproducing the HF component of the spectral decomposition, less for the LF part. Considering our findings about the planetary disturbances non-linearity a question rises: How the non-linear behaviour, previously depicted, could be reproduced by state of the arts coupled GCMs? This question is of paramount importance if we want trust the statistical description of our near future performed by the climate models (statistical extremes etc.).

Considering the “best” climate models of the previous sub-section, in terms of their skill in reproducing the statistical properties of mid-latitude atmospheric waves, we could answer this question at least for our best choice. Only for two models – GFDL-CM2.1 and MIROC(hires) - the climatologies of the wave activity are statistically consistent with that of the reanalyses, both for the global and process-oriented metrics. The fact that these two models are the only two performing well in both metrics suggest that they overperform the other models in the representation of the mid-latitde winter atmospheric dynamics of the NH. The span of performances of the considered set of climate models in the representation of the tropospheric waves is a rather relevant issue, also because all models are expected to agree reasonably in other diagnostics such as the average vertical temperature profiles, average surface temperature, precipitation and so on (see e.g. the results presented in the Third Assessment Report of the IPCC).

Considering the planetary wave index (WAI), computed using the wind field by the mean of the spectral properties of the space derivative, and the jet strength index (JSI) as for the observational data-sets, we compute the 2-D pdf for the two aforementioned models (Figure 20.8). Several interesting features appear: 1. both the models show some non-linear features in their 2-D pdf plots; 2. the GFDL model has the “correct”, with respect to the re-analysis, WAI range of values, while the JSI is

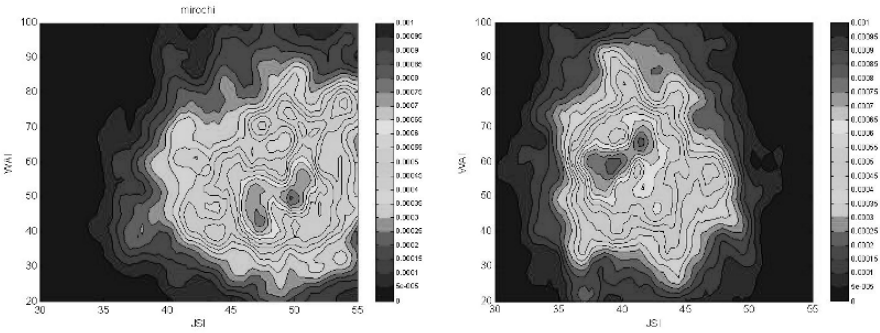


Fig. 20.8. Evidence of the dependence of the planetary wave regimes on the jet strength for the winter-time Northern Hemisphere. Illustrated is the two-dimensional joint pdf using the index of the planetary waves (WAI) and the index of the tropospheric jet strength (JSI). GFDL and MIROCHR models, present-climate.

shifted toward lower values; 3. the MIROC model presents a shift towards high values for the JSI index, and towards low values for the WAI.

5 Conclusions

Since synoptic and planetary waves provide a relevant contribution to the meridional transport of energy and momentum in the mid-latitudes, the physical processes driving such phenomena are relevantingredients for the characterization of the climate system (Speranza, 1983; Peixoto and Oort, 1992), and discrepancies may be critical when interpreting the results of unconstrained climate models, possibly reducing the significance of the resulting simulations of past and future climatic conditions. In fact, it was confirmed in a number of studies that global climate models on the average featured a large overestimation of the baroclinic synoptic waves (e.g. Sumi and Kanamitsu, 1984; Klinker and Capaldo 1986; Siegmund 1995) and a large underestimation of the planetary component (e.g. Wallace et al., 1983; D'Andrea et al., 1998), with biases sometimes of the same size as the average climatology of the observations and analyzed atmospheric fields. Since the 80's, the theory on the dynamical interaction between the low and high frequency disturbances (Haines, 1994) could be advocated for explaining these models biases. Nevertheless, this approach could underestimate the non-linear processes acting at the planetary scale and accounting for a large part of the bias in the low frequency domain.

A theory was developed to account for the observed multimodal (or, anyway, non-normal) distribution of the activity of planetary diturbances, which constitute the low frequency component of the atmospheric variability. The planetary disturbances were soon assumed to be related to orographic resonance processes (Charney and Devore, 1979), and extended including the non-linear self interaction which features fixed strength of the westerlies as observed (Benzi et al., 1986). In this chapter and in Ruti et al. (2006), we revise this theory by analyzing both planetary waves index and jet strength index.

We show that the upper tropospheric jet strength is a critical parameter in determining whether the planetary waves indicator exhibits a uni- or bimodal behavior. These results are obtained by considering the data of the NCEP-NCAR and ECMWF reanalyses for the overlapping period. For strong jet strength, the planetary waves depict a bimodal behavior, while reducing the jet strength their pdf is unimodal. So, in the Northern Hemisphere, the atmospheric phase-space can be characterized in terms of the sub-tropical jet, consistent with the physical framework proposing the nonlinear resonance folding of the topographic resonance as the basic mechanism for the low frequency variability of the atmosphere.

The ability of climate models in reproducing the non-linear waves self interaction has been analyzed by using the simulations of the state-of-the-art climate models gathered by the Project for Climate Model Diagnostics and Intercomparison (PCMDI: http://www-pcmdi.llnl.gov). First, we analyze the propagating and standing component in the observed and simulated atmosphere. Computing the Hayashi spectra of the 500hPa geopotential height fields, we inspect the variability observed in the Northern Hemisphere on the high frequency and low frequency spectral subdomains. Taking into account the two realization of the atmospheric circulation for the period 1961-2000 (the two reanalysis), our results show that, considering the HF and LF spectral features of the atmosphere, the baroclinic waves are typically overestimated by the climate models, while the planetary waves are usually underestimated. This closely resembles the results of many diagnostic studies performed in the past on global weather forecasting models (Tibaldi 1986), but continues to suggest a link between the statistical properties of the low frequency variability and the high frequency variabiliy.

Only for two models – GFDL-CM2.1 and MIROC(hires) - the climatologies of the spectral activity are statistically consistent with that of the reanalyses. The fact that these two models are the only two performing well in both metrics suggest that they overperform the other models in the representation of the mid-latitude winter atmospheric dynamics of the NH. Nevertheless, the analysis of the planetary waves index and of the jet strength index, in terms of bi-dimensional pdf, shows that the models have still problems in reproducing the right non-linear behaviour of the planetary disturbances. It would be a relevant step to use these models for sensitivity runs in order to characterize their parameter-space.

Acknowledgments

The authors wish to thank A. Sutera for useful suggestions. NCEP data have been provided by the NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, from their web site at http://www.cdc.noaa.gov/. The ECMWF ERA-40 data have been obtained from the ECMWF data server at http://data.ecmwf.int/data/. This work is based on a s-meeting held at the University of Camerino in July 2004. A s-meeting means few days workshop focused on 2-3 topics without any time limitation. All the authors contributed equally to this work and to the s-meeting. All the authors have equally contributed to this chapter.

References

Annamalai H., Slingo J. M., Sperber K. R., Hodges K., (1999) The mean evolution and variability of the asian summer monsoon: comparison of ECMWF and NCEP-NCAR reanalyses. Mon. Wea. Rev., 127, 1157-1186.

Baur, F. (1951), Extended range weather forecasting. Compendium of Meteorology, Amer. Meteorol. Soc., 814-833.

Benzi R, Malguzzi P, Speranza A Sutera A (1986) The statistical properties of general atmospheric circulation: observational evidence and a minimal theory of bimodality. Q J Roy Met Soc 112: 661-674

Benzi R, Speranza A (1989) Statistical properties of low frequency variability in the Northern Hemisphere. J Clim 2: 367-379

Blackmon ML (1976) A climatological spectral study of the 500 mb geopotential height of the Northern Hemisphere. J Atmos Sci 33: 1607-1623

Buzzi A, Trevisan A, Speranza A (1984) Instabilities of a baroclinic flow related to topographic forcing. J Atmos Sci 41: 637-650

Caires, S., Sterl, A., Bidlot, J.-R., Graham, A., and Swail, V., (2004) Intercomparison of different wind wave reanalyses, J. Climate 17, 1893-1913.

Charney JG, DeVore JG (1979) Multiple flow equilibria in the atmosphere and blocking. J. Atmos Sci 36: 1205-1216

Charney JG, Straus DM (1980) Form-drag instability, multiple equilibria and propagating planetary waves in the baroclinic, orographically forced, planetary wave system. J Atmos Sci 37: 1157-1176

Christiansen, Bo, (2005) On the bimodality of planetary-scale atmospheric wave amplitude index. J. Atmos. Sci.,. In press.

Corti, S., Molteni F., Palmer, T.N. (1999) Signature of recent climate change in frequencies of natural atmospheric circulation regimes. Nature, 398, 799-802.

D'Andrea, F., S. Tibaldi , M. Blackburn, G. Boer, M. Deque, M. R. Dix, B. Dugas, L. Ferranti, T. Iwasaki, A. Kitoh, V.Pope, D. Randall, E. Roeckner, D. Straus, W. Stern, H. Van den Dool and D. Williamson, (1998), Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979–1988. Climate Dynamics, 14, 385-407

Dell'Aquila, A., Lucarini, V., Ruti, P.M., and S. Calmanti (2005) Hayashi spectra of the northern hemisphere mid-latitude atmospheric variability in the NCEP-NCAR and ECMWF reanalyses. Climate Dynamics DOI: 10.1007/s00382-005-0048-x.

Dell'Aquila, A., Ruti P.M., and A. Sutera, (2006) Effects of the baroclinic adjustment on the tropopause in the NCEP–NCAR reanalysis. Clim. Dyn. DOI 10.1007/s00382-006-0199-4

Dole, R. M. (1983) Persistent anomalies of the extratropical Northern Hemisphere wintertime circulation. Large-Scale Dynamical Processes in the Atmosphere, B. J. Hoskins and R. P. Pearce, eds., Academic Press, NY, 95-109.

Fasano, G. and Franceschini, A. (1987) Mon. Not. R. Astr. Soc. 225, 155–170.

Fraedrich K, Bottger H. (1978) A wavenumber frequency analysis of the 500 mb geopotential at 50°N. J. Atmos. Sci. 35, 745-750

Haines, K., (1994) Low frequency variability in atmospheric middle latitudes. Surveys in Geophysics, 15, 1-61.

Hansen, A. R. and A. Sutera (1986) On the probability density distribution of planetary-scale atmospheric wave amplitude. J. Atmos. Sci. 43, 3250--3265.

Hansen, A. R., Sutera A. and D. E. Venne (1989) An examination of midlatitude power spectra: evidence for standing variance and the signature of El Niño. Tellus. 41(A), 371-384.

Hayashi Y (1971) A generalized method for resolving disturbances into progressive and retrogressive waves by space Fourier and time cross-spectral analysis. J Meteorol Soc Jap 49: 125-128

Hayashi Y (1979) A generalized method for resolving transient disturbances into standing and travelling waves by space-time spectral analysis. J Atmos Sci 36: 1017-1029

Harrison M. J., Rosati A., Soden B. J., Galanti E., Tziperman E., (2002) An evaluation of Air-Sea Flux Products for ENSO Simulation and Prediction. Mon. Wea. Rev. 130, 723-732

Hodges K. I., Hoskins B. J., Boyle J., Thorncroft C., (2003) A comparison of recent reanalysis datasets using objective feature tracking: storm tracks and tropical easterly waves, Mon. Wea. Rev. 131, 2012-2037.

Josey S. A., (2001) A comparison of ECMWF, NCEP-NCAR, and SOC surface heat fluxes with moored buoy measurements in the subduction region of the North Atlantic. J. Climate, 14, 1780-1789.

Josey S. A., Kent E. C., Taylor P. K., (2002) Wind stress forcing of the ocean in the SOC climatology: comparison with the NCEP-NCAR, ECMWF, UWM/COADS, and Hellerman Rosenstein datasets. J. Phys. Oceangr., 32, 1993-2019

Kistler R, et al. (2001) The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. Bull. Am. Meteorol. Soc. 82, 247–267

Klinker, E. and M. Capaldo, (1986) Systematic errors in the baroclinic waves of the ECMWF model. Tellus, 38A, 215-235

Landau, L. D. and Lifschitz, E. M., (1976) Mechanics – Course of Theoretical Physics, 3rd ed. Pergamon Press, Oxford, pp. 170

Landau, L. D. and Lifschitz, E. M., (1980) Statistical Physics – Course of Theoretical Physics, 3rd ed. Pergamon Press, Oxford, pp. 544

Lucarini, V., S. Calmanti, A. Dell'Aquila, PM Ruti, and A. Speranza, (2006) Intercomparison of the northern hemisphere winter mid-latitude atmospheric variability of the IPCC models. Clim. Dyn., In press.

Malguzzi, P. and Speranza, A. (1981) Local Multiple Equilibria and Regional Atmospheric Blocking. J. Atmos. Sci. 9: 1939–1948.

May W (1999) Space-time spectra of the atmospheric intraseasonal variability in the extratropics and their dependency on the El Niño/Southern Oscillation phenomenon: model versus observation Climate Dynamics (1999) 15: 369-387

Peacock, J.A. (1983) Mon. Not. R. Astr. Soc. 202, 615–627.

Peixoto, J.P. and Oort, A.H. (1992). Physics of Climate, Chapter 1, American Institute of Physics, New York, USA.

Pratt, R. W., (1976) The interpretation of space-time spectral quantities. J. Atmos. Sci., 33, 1060–1066.

Renfew I. A., Moore G. W. K., Guest P. S., Bumke K., (2002) A comparison of surface layer and surface turbulent flux observation over the Labrador Sea with ECMWF Analyses and NCEP Reanalyses. J. Phys. Oceanogr., 32, 384-400.

Rex, D.F. (1950) Blocking action in the middle troposphere and its effect upon regional climate. Part 2: The climatology of blocking action. Tellus, 2, 275-301.

Ruiz-Barradas, A., S. Nigam, (2004) Warm-season rainfall variability over the US Great Plains in observations, NCEP and ERA-40 reanalyses, and NCAR and NASA atmospheric model simulations, J. Climate, in press

Ruti, P. M., V. Lucarini, A. Dell'Aquila, S. Calmanti, and A. Speranza (2006), Does the subtropical jet catalyze the midlatitude atmospheric regimes?, Geophys. Res. Lett., 33, L06814, doi:10.1029/2005GL024620

Siegmund P. (1995) The generation of available potential energy: a comparison of results from a general circulation model with observations. Climate Dynamics, 11, 129-140.

Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis. Chapman & Hall.

Simmons, A. J., Gibson, J.K. (2000) The ERA-40 Project Plan, ERA-40 Project Report Series No. 1, ECMWF, 62 pp.

Speranza A., (1983) Deterministic and statistical properties of the westerlies. Paleogeophysics 121: 511-562

Stephenson, D.B., Hannachi, A., O'Neill, A. (2004) On the existence of multiple climate regimes. Quart. J. Roy. Mel. Soc., 130, 583-605.

Sturaro, G., (2003) A closer look at the climatological discontinuities present in the NCEP/NCAR reanalysis temperature due to the introduction of satellite data. Climate Dyn., 21, doi:10.1007/s00382-003-0348-y.

Sumi A. and M. Kanamitsu (1984) A study of systematic errors in a numerical prediction model, Part I: General aspects of the systematic errors and their relation with the transient eddies. J. Met. Soc. Japan 62: 234-251.

Tibaldi, S. (1986) Envelope orography and maintenance of the quasi-stationary circulation in the ECMWF global models, Advances in Geophysics, 29, pp. 339-373.

Van Oldenburg G.J., S. Philip, M. Collins (2005) El Niño in a changing climate: a multi-model study, Ocean Science Discussions, 2, 267-298 (2005)

Wallace, J.M., S. Tibaldi and A. J. Simmons (1983) Reduction of systematic forecast errors in the ECMWF model through the introduction of an envelope orography. Quart. J. R. Met. Soc., 109, 683-717.

Wallace J.M., Lim G.H., Blackmon M.L. (1988) Relationship between cyclone tracks, anticyclone tracks and baroclinic waveguides. J Atmos Sci 45: 439-462

21 On the Spatiotemporal Variability of the Temperature Anomaly Field

Leila M. V. Carvalho[1], Anastasios A. Tsonis[2], and Charles Jones[3]

[1] University of São Paulo, Dept. of Atmospheric Sciences, leila@icess.ucsb.edu
[2] University of Wisconsin Milwaukee, Dept. of Mathematical Sciences, aatsonis@uwm.edu
[3] University of California Santa Barbara, ICESS, cjones@icess.ucsb.edu

Abstract. In this study, low-frequency variations in temperature anomaly are investigated by mapping temperature anomaly records onto random walks. We show evidence that global overturns in trends of temperature anomalies occur on decadal time-scales as part of the natural variability of the climate system. Paleoclimatic summer records in Europe and New-Zealand provide further support for these findings as they indicate that anti-persistence of temperature anomalies on decadal time-scale have occurred in the last 256 yrs. Atmospheric processes in the subtropics and mid-latitudes of the SH and interactions with the Southern Oceans seem to play an important role to moderate global variations of temperature on decadal time-scales.

1 Introduction

The Earth's average surface temperature has increased by 0.6 ± 0.2°C since the late 19th century with significant global impacts. The dramatic heat waves in the Northern Hemisphere (Berniston 2004; Karl, Knight, and Plummer 1995; Karl and Knight 1997; Meehl and Tebaldi 2004), extensive collapses of ice shelves over the Artic (Serreze, Maslanik, Scambos, Fetterer, Stroeve, Knowles, Fowler, Drobot, Barry, and Haran 2003) and Antarctic (Shepherd, Wingham, Payne, Skvarca 2003), recent hurricane activity (Emanuel 2005) and other major weather events are viewed as a direct result of this warming, which is attributed partly to natural variability of the climate system and partly to human activity. However, uncertainties about the details of the natural variability as well as the response of the climate system to the rapid increase of greenhouse gases still remain (IPCC 2001). Understanding natural variability of the climate system is crucial to predict non-linear climate forcings and and abrupt changes possibly due to anthropogenic activity (Alley, Marotzke, Nordhaus, Overpeck, Peteet, Pielke Jr., Pierrehumbert, Rhines, Stocker, Talley, and J. M. Wallace 2003). The objective of this study is to improve our understanding of climate variability by providing new insights about the spatiotemporal variability of temperature anomalies.

Here, low-frequency variations in temperature fluctuations on distinct time-scales are investigated by mapping temperature anomaly records onto random walks (RW) (Tsonis 1992; Tsonis, Hunt, and Elsner 2003). RW based statistics of temperature fluctuations have been successfully applied to demonstrate the importance of El Niño in moderating changes in global temperature (Tsonis et al. 2003). Motivated by this previous study, RW is used here as a powerful time series approach to describe low-frequency variations in the sequence of anomalous events over time in a given region and ultimately understand temporal variations and spatial organization of global temperature fluctuations.

This chapter is organized as follows: section 2 discusses the data and the rationale for the application of RW to interpret low-frequency variations of global temperature fluctuations. Section 3 uses RW correlation and scaling properties to investigate the role of low-frequency variations in the atmosphere in moderating global temperature fluctuations. Section 4 discusses persistence and anti-persistence in trends of temperature fluctuations on decadal time-scales. Section 5 presents an analysis of paleoclimatic data, which confirms this anti-persistence on decadal time-scales in the last 250 years. Section 6 summarizes the main results and conclusions of this study.

2 Data and Methods

Global daily temperature anomalies are examined in the present climate (1948-2005) using NCEP/NCAR reanalysis (Kalnay et al. 1996) and over the past 225 yrs (1750-1975) using reconstructed summer temperature over Europe (Briffa, Jones and Schweingruber 1988) and New Zealand (Xiong and Palmer 2000). Long-term trends are removed from daily temperatures at 2m (T2m) in the present climate and anomalies T'(t) are obtained by further removing the annual and semi-annual cycles.

The global spatiotemporal variations of T'(t) are interpreted within the framework of random-walks (RW). According to the RW approach (Tsonis et al. 2003; Tsonis 1992), the corresponding $T'(t)$ at a given location is mapped onto a walk by calculating a net displacement $y(t)$, defined by the running sum:

$$RW \equiv y(t) = \sum_{i=1}^{t} T_i'(t) \tag{1}$$

The RW describes the sequence of anomalous events over time in a given region and is influenced by the magnitude and persistence of the fluctuations; these, in turn, depend on complex relationships with the climate system (Tsonis 1992, Mann, Bradley and Hughes 1998). RW is an efficient approach to represent important low-frequency variations in the regime of temperature fluctuations (Tsonis et al. 2003). Even though in the limit $t \to \infty$ a random walk is stationary, over finite time scales it exhibits nonstationarity. Thus, if climate forcings in a period Δt are such that positive (negative) T'(t) are more frequent than negative (positive) T'(t), the corresponding RW exhibits a positive (negative) slope in that time interval. Therefore, time variations in the slope of RW are important indications of regime changes in temperature fluctuations.

We illustrate these properties by showing the time evolution of y(t) (Eq. 1) for a periodic time-series constructed with the same number of data points as the observations (Fig. 21.1). A periodic time-series (in this case, sin($4\pi/n$), with n=21170) was used because it is stationary and cycles can be interpreted as idealized periodic changes in regimes identified by the sign of the anomalies. The application of Eq. 1 to a periodic time-series yields periodic walks, as illustrated by the dashed line in Fig. 21.1. Nevertheless, these periodic walks clearly show that changes in regime correspond to changes in the slope of RW. Local minima and maxima in the RW indicate the time when changes in regime occur. This useful property of Eq. 1 can be used to interpret low-frequency variations in *T'(t)*.

Another appropriate statistical quantity used to characterize the walk is the root mean square fluctuation about the average displacement (Tsonis et al. 2003):

$$F(t)=[<[\Delta y(t)]^2> - <[\Delta y(t)]>^2]^{1/2} \quad (2)$$

where $\Delta y(t) = y(to + t) - y(to)$, and the symbols (< >) indicate average over all positions of the walk. When $F(t) \propto t^H$ it can be shown that there are three possible types of behavior: (1) uncorrelated time series with H=0.5; (2) time series exhibiting positive long-range correlations with H> 0.5 and (3) time series exhibiting negative long range correlations with H<0.5. The importance of H coefficient is that the correlation function $C(t)$ of $y(t)$ with past elements $y(-t)$ is given by $C(t)=2(2^{2H-1}-1)$ (Feder 1988). The implications of this relationship is that for H=0.5, $C(t)$ is independent of t. For H>0.5, $C(t)$ is positive, which means that trends in fluctuations in the future are

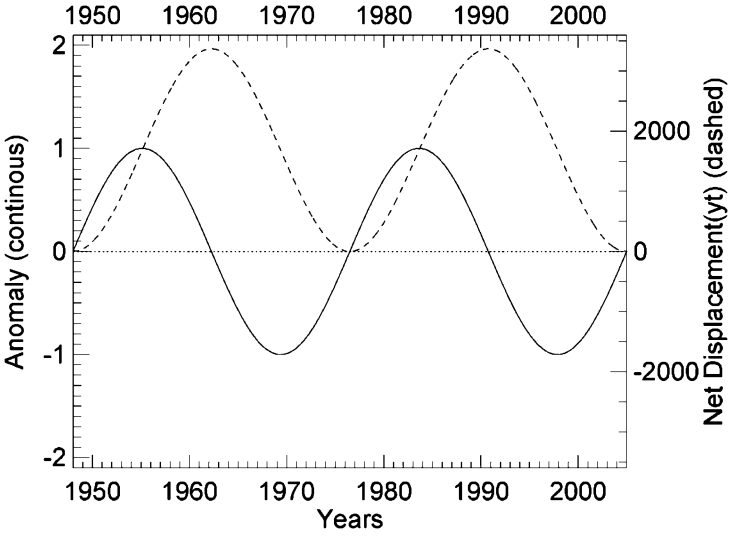


Fig. 21.1. Example of RW (dashed line) obtained from Eq(1) applied to a periodic time series *sin*(4π/n), with n=21170 (solid line).

positively correlated with trends in fluctuations in the past. On the other hand, for $H<0.5$, $C(t)$ is negative implying that decreasing (increasing) trends in the past are 60 followed by increasing (decreasing) trends in fluctuations in the future. This concept differs from the ordinary idea of persistence of a given pattern in time, but reflects the existence of scale invariance, a property inherent to many nonlinear processes in nature (Tsonis 1992). Therefore, persistence is used here to indicate the tendency for existing temperature fluctuations regime to persist. It is reflected by the positive correlation between successive values of anomalies in a given time-scale. These relationships are used here to examine low-frequency variations in global temperature.

3 Time-Correlation Scaling Properties

Tsonis, Roebber, and Elsner (1998) applied the RW approach to monthly global temperature data (1865-1998). They showed that the global temperature data exhibit two distinct scaling regimes; one consistent with a power law with $H=0.65$ for $t<20$ months and the other with a power law with $H<0.4$ for $t>20$ months. The interpretation is that processes of time scales less than 20 months sustain a tendency toward an initial trend (whether positive or negative) and processes of time-scales greater than 20 months tend to reverse the past trend. They also conclude that this change in scale defines an important characteristic time scale in global climate. Here we extend this type of analysis to investigate the spatial properties of scaling of temperature fluctuations.

In order to examine the spatial variation of scaling properties, we look at variations in the power-law (H coefficient) in pre-defined scales that are known to explain a large fraction of spectral variance of temperature and circulation over the globe between 10^{-3}– 7 years (Peixoto and Oort 1992; Madden and Julian 1994). This interval of scales yields robust statistics for the determination of H, considering the length of data. The scales we investigate correspond to the following time-intervals: 2–15 days, 20–90 days, 100–200 days, 2–4 years, 4.5–6 years, and 6.5–7 years. The separation in scales does not imply that we expect a break in scaling properties on these particular scales. We emphasize that linear correlations obtained from $\log(F(t))$ vs. $\log(t)$ are above 0.99 for all scales (not shown).The spatial variability of H for the different time-scales is shown in Fig. 21.2.

On synoptic time-scales (2-15 days) (Fig. 21.2a), H coefficient is above 0.5 over all regions of the planet which indicates a universal persistence. H is higher over tropical Pacific and Atlantic ($H>0.90$). The subtropics and extratropics, particularly over the storm track of the Southern Hemisphere and eastern coast of the United States, Southern Australia and southern South America, are associated with relatively lower values of H, between 0.6 and 0.7. Cold fronts, cyclones and anticyclones are some of the known phenomena affecting temperature on the synoptic time-scales and they occur in nature essentially to equilibrate differences in temperature between tropics and extratropics (e.g. Holton 2004). Nevertheless, although changes in temperature are certainly observed before and after a passage of a frontal system, these changes do not reverse the long-term trends in temperature anomalies.

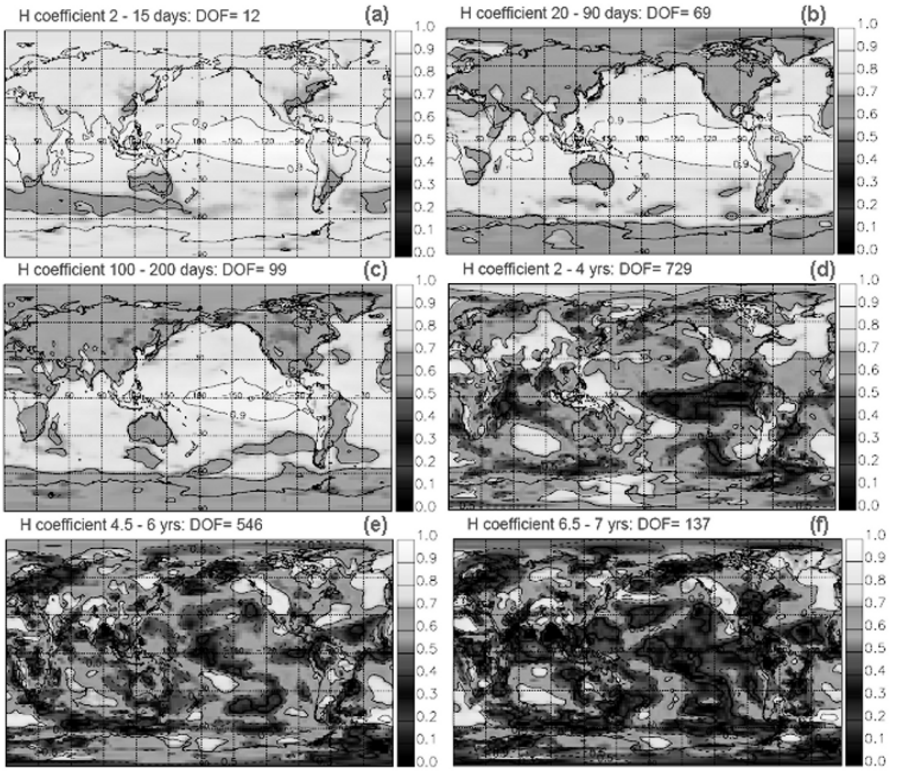


Fig. 21.2. The scaling coefficient *H* obtained in distinct time-scales: 2-15 days (a), 20-90 days (b), 100-200 days (c), 2-4 years (d), 4.5-6 years (e) and 6.5-7 years(f). Degrees of freedom are indicated at the top of each panel. Dashed (Solid) lines identify $H<0.5$ ($H>0.5$). (A color version of this figure appears between pages 196 and 197).

On intraseasonal time-scales (20-90 days) (Fig. 21.2b), the universal tendency for persistence continues albeit with lower values ($0.55 < H < 0.65$) over continental areas. The Madden-Julian oscillation (Madden and Julian, 1994) is one of the most powerful tropical phenomenon on intraseasonal time-scales that modulates cloudiness, precipitation and circulation over the tropics, with impacts on the subtropics and extratropics of both hemispheres (e.g, Lau and Waliser 2005 and references therein). In addition, important variations on intraseasonal time-scales have been observed in temperature (Przybylak 2002) and sea ice (Baba, Minobe, Kimura and Wakatsuchi 2006) records in high latitudes. However, the physical reasons for distinct scaling characteristics and, therefore, *H* over the continents and oceans are not obvious. What these results suggest, however, is that temperature fluctuations over oceans and lands respond distinctly to perturbations on intraseasonal time-scales.

The 100-200 days time-scale (Fig. 21.2c) shows spatial characteristics that are similar to intraseasonal (20-90 days) time-scales, in particular low *H* values over the

continents. Nevertheless, H decreases in most regions and now approaches to 0.50, which indicates that for this time-scale future and past RW trends are likely uncorrelated. In some few areas over the extratropics of the Northern Hemisphere H is < 0.50. In addition, changes in *H* also occur over the southern oceans on this scale.

Finally, we focus our attention on scaling properties on interannual to intra-decadal time scales (2-7 years) (Figs. 21.2d-f). The general observation now is that as the time scale increases the tendency for *H* is to become less than 0.5 almost everywhere on the globe. Arguably, these time scales correspond to the time scale of ENSO, one of the most important features of our climate system. This observation supports the hypothesis postulated by Tsonis et al. (2003) that El Nino is a mechanism that reverses previous positive trends of temperature anomalies.

Therefore, while for very short time scales processes in the climate system tend to promote an existing trend, for longer time scales processes are activated to reverse an existing trend and to establish stationarity; consistent with the results obtained for the global temperature (Tsonis et al. 1998). Thus, the conclusions reached by using the global temperature record are not just the result of some averaging procedure but reflect the properties of the small scale fluctuations in the temperature field. Due to insufficient statistics in estimating *F(t)* at longer time scales we cannot extend this analysis any further in time, but a different approach indicates that this anti-persistence holds for at longer time scales as well.

4 RW Anti-Persistence on Long Time-Scales

Remarkable changes in the slope of RW have occurred on decadal time-scale in large areas over the globe during 1948-2005. Some examples are shown over selected regions in the extratropics of the Northern (Fig. 21.3a) and Southern (Fig. 21.3b) Hemispheres. Decadal variations in the RW are clearly observed. In addition, inter-annual to intra-decadal variations are also evident, particularly for the Northern Hemisphere (Fig. 21.3a). The spatial coherence of RW is also obvious from the similar temporal variations of RW among neighbor grid points (Fig. 21.3a,b) .

However, given the high spatial correlation observed among neighbour RW and the non-linear nature of temperature fluctuations, a question arises: can surrogate time-series with the same lag one auto-correlation yield similar decadal variability by chance? In order to examine this issue we produce Markov processes with the same length, and lag one autocorrelations as the original T2m time series of Fig. 21.3a and repeat the analyses. Figure 21.4 clearly indicates that decadal variations similar to the actual data (Fig. 21.3a) are unlikely occurring by chance or just by preserving the short term memory and the spatial pattern of variability (indicated by the standard deviation) of the original time-series. RW from surrogated data in general exhibits variations in slope on short time-scales (Fig. 21.4).

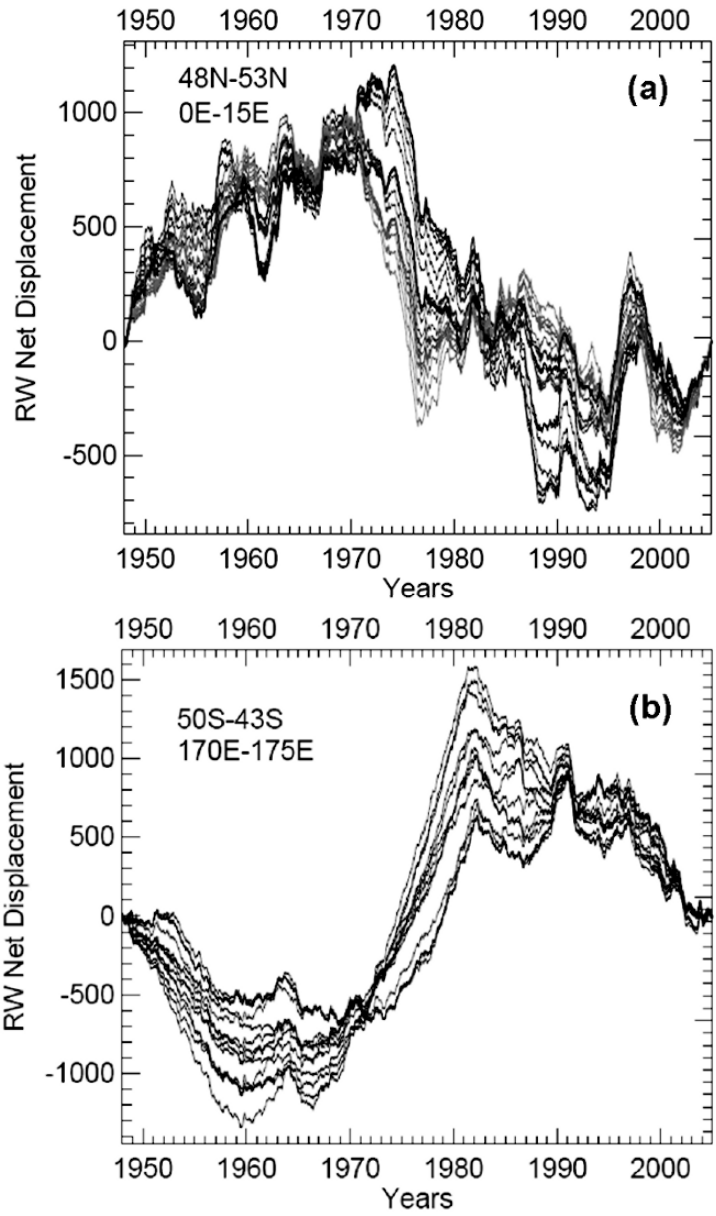


Fig. 21.3. Random walks computed for daily temperature anomalies in midlatitudes of the Northern and Southern Hemispheres. Random walks (1948-2005) observed over Northern Europe (48-53N, 0E-15E) (a). Random walks observed in the same period over South Pacific around New Zealand (40-55S, 170-178E) (b). Each RW curve is obtained for a distinct grid point inside the domain. The ordinates show the daily net displacement y(t) (Eq.1) and the abscissa indicates time.

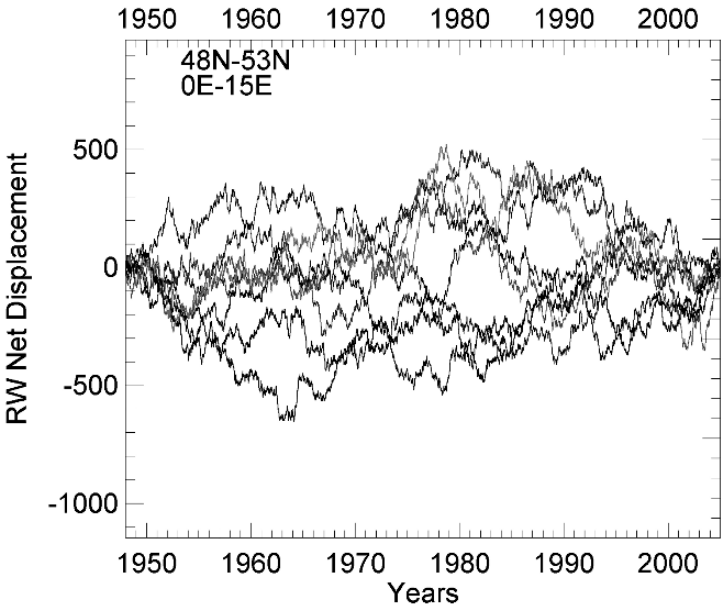


Fig. 21.4. Random walks computed for surrogate Markov lag-1 processes for daily temperature anomalies in midlatitudes over Northern Europe (48-53N, 0E-15E). Each RW curve is obtained for a distinct grid point in the domain.

Therefore, the almost symmetrical decadal variability of the random walks over regions in the Northern (Fig. 21.3a) and Southern (Fig. 21.3b) Hemispheres suggest the existence of global anti-persistence of temperature fluctuations on decadal time-scales.

Over finite time scales, RW exhibits nonstationarity. Therefore, the total RW length (RWL), defined as $\sum y(t)$ over a given period of years is a useful quantity which may yield further evidence of the anti-persistence in temperature anomalies on long time-scales. In order to investigate this problem, random walks were arbitrarily computed in two non-overlapping periods: 1948-1975 (Fig. 21.5a) and 1976-2005 (Fig. 21.5b).

Interestingly, areas with high RWL magnitude (positive and negative) switch sign from the 1948-1975 to the 1976-2005 in a coherent and consistent way with maxima in Fig. 21.5a becoming minima in Fig. 21.5b and vice-versa. We observe a general tendency in the first period (Fig. 21.5a) for positive RWL values in the northern hemisphere and Antarctica (region 1) and a general tendency for negative RWL values in oceans of the southern hemisphere (region 2). This is an indication that in region 1, during 1948-1975, the regime of temperature anomalies was such that the frequency and/or intensity of positive temperature anomalies in those years, in contrast with negative anomalies, were such that positive trend in the random walk is observed, whereas in region 2 the trends are negative.

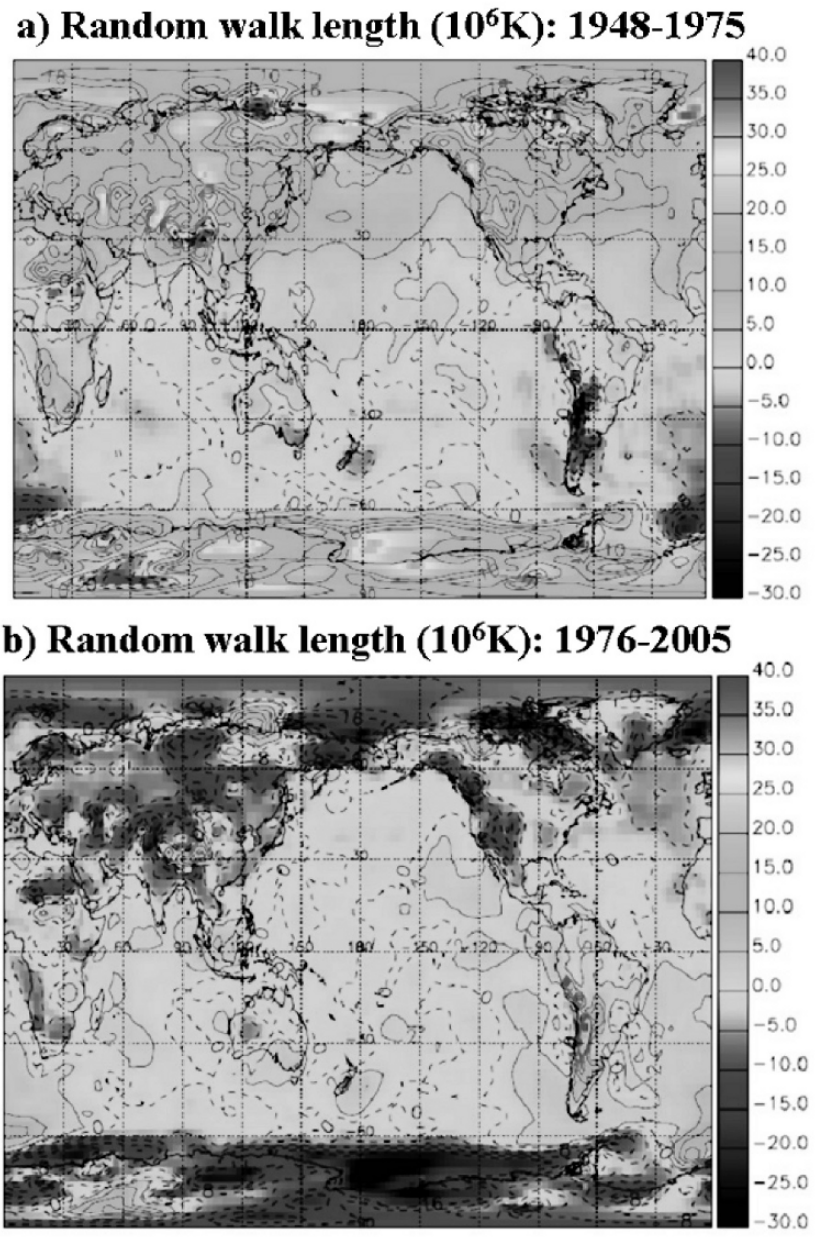


Fig. 21.5. Random walk length (RWL) 1948-1975 (a) and 1976-2006 (b). Solid (dashed) lines enclose regions with positive (negative) RWL. Units are 10^6 K Note that maxima in 5a become minima in 5b and vice-versa. (A color version of this figure appears between pages 196 and 197).

This is an interesting observation but what is remarkable is the reversal observed in Fig. 21.5b. Now the RWL is negative in region 1 and positive in region 2. The reversal appears to be universal; it is occurring at almost all grid points. This reversal suggests that anti-persistence in temperature fluctuations extends to time scales longer than those of ENSO. This will indicate that decadal time-scales processes are also important in regulating temperature variations.

5 Anti-persistence in Paleoclimatic Records

Next, we verify the anti-persistence at decadal time scales using paleoclimate data. Fortunately, temperature records exist for Europe (European Gridded and Regional summer reconstruction) (Briffa et al. 1988) and New Zealand (Xiong and Palmer 2000), where as Figs. 21.3 and 21.5 show when RW in Northern Europe is positive

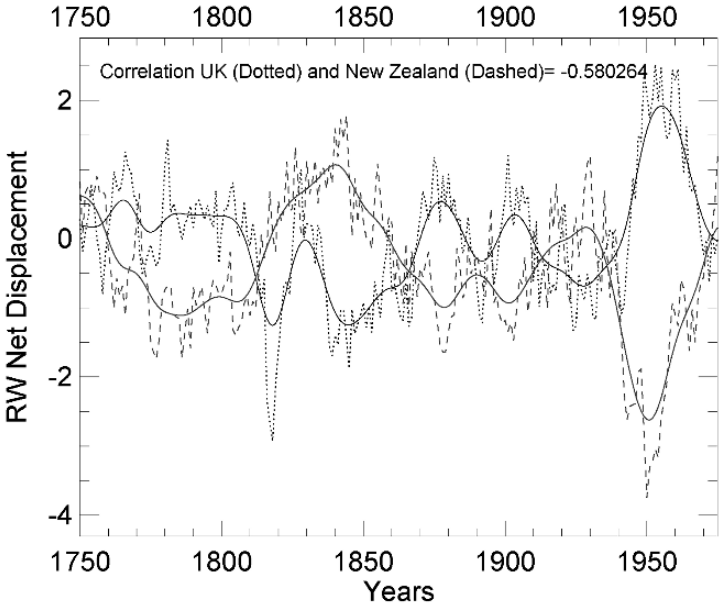


Fig. 21.6. RW of summer reconstructed temperature anomalies in New Zealand (dashed line) and United Kingdom (dotted line) during 1750-1975. Correlation between "raw" RW of reconstructed paleoclimatic data is equal -0.58 and is statistically significant at 5% significance level. The number of degrees of freedom (N_{dof}) is equal to 15 based on the relationship $N_{dof}=N(1-\rho_{1NZ}\,\rho_{1UK})/(1+\rho_{1NZ}\,\rho_{1UK})$, where N is the number of seasons and ρ_{1NZ} (ρ_{1UK}) is the lag-1 auto-correlation function for New Zealand (United Kingdom) records. Solid lines are 5 years moving average (200 passes) to highlight low-frequency variability of temperature regimes in New Zealand and United Kingdom.

around New Zealand is negative and vice versa. Temperature variations over Europe (EU) are reconstructed for a 6-month summer season (April-September) for the period 1750-1975 using a network of maximum latewood (Briffa et al. 1988). The New Zealand (NZ) data consist of average summer (February-March) temperature derived from eleven NZ cedar ring width chronologies (Xiong and Palmer 2000).

Temperature anomalies in both datasets are obtained with respect to the 1951-1970 period. Trends are removed and the random walks of summer temperature fluctuations are computed according to Eq.1. Figure 21.6 shows the two random walks (dotted lines). The correlation coefficient between these two random walks is -0.58 and is statistically significant at the 5% significance level. The solid lines are 5-year moving average and show correlation equal -0.78. These results indicate that the remarkable anti-persistence on decadal time-scale and the anti-correlation between New Zealand and UK datasets have been present in the last 2.5 centuries. The reconstructed datasets are summer averages in both hemispheres and, consequently, the anti-correlation is not an artefact of computing random walks over distinct seasons. The slopes in the RW indicate changes in regimes of seasonal temperature fluctuations. The anti-correlation provides additional support to the results shown in Figs. 21.3 and 21.5 that suggest that atmospheric processes in the subtropics and mid-latitudes of the SH and interactions with the southern Oceans play an important role to moderate global variations of temperature on decadal time-scales.

Note that the scaling properties (Fig. 21.7) obtained from RW from New Zealand and United Kingdom are very consistent with those in Fig. 21.5. The scaling coefficient H for

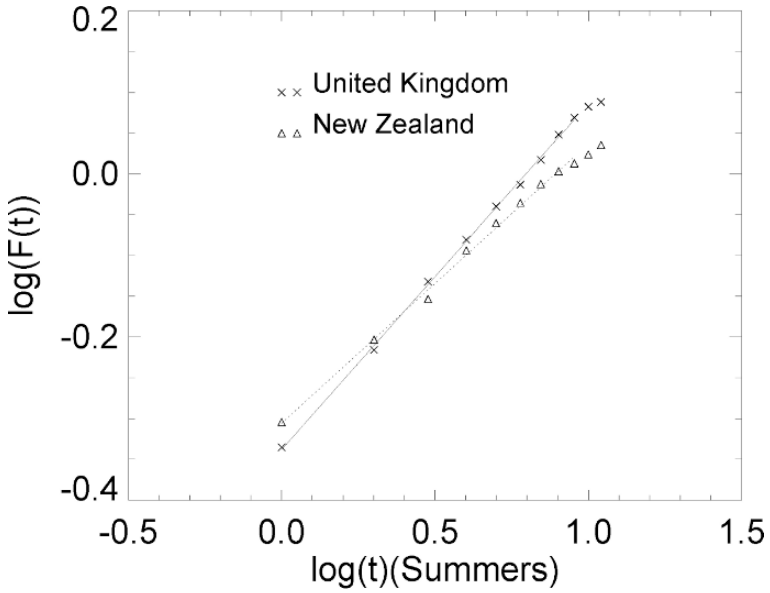


Fig. 21.7. Scaling law for Paleoclimatic RW (1-11 years) at United Kingdom (solid line) and New Zealand (dotted line).

United Kingdom is equal to 0.42 and for New Zealand is equal to 0.34 within the 1-9 years time scale range. Similar *H* coefficients are observed around New Zealand and Northern Europe typically at 6-7.5 years time-scale.

6 Conclusions

The anti-persistence of the temperature filed on inter-decadal time scales is part of the decadal variability of the climate system and this property has not been identified before. Processes at longer time scales than that of ENSO are also responsible for maintaining stationarity in the temperature anomaly field. The origin of interdecadal fluctuations in the climate system is one of the most challenging problems in climate dynamics. For example, the interdecadal fluctuations of the large-scale circulation over the North Pacific (Deser, Phillips and Hurrell 2004; Trenberth and Hurrell 1994) has been hypothesized to occur from a variety of possibilities ranging from coupled ocean-atmosphere interactions (Latif and Barnett 2005), to air-sea interactions in the subtropical Pacific with atmospheric teleconnections to the North Pacific (Knutson and Manabe 1998), to extratropical-tropical interactions via an equatorward subsurface oceanic pathway and a return atmospheric bridge (Gu and Philander, 1997), to climate noise arising from integrating stochastic atmospheric variability over the North Pacific Ocean mixed layer and the thermocline (Pierce, Barnett, Schneider, Saravanan, Dommenget and Latif., 2001; Frankignoul, Kestenare, Sennechael, Coetlogon and D'Andrea 2000). Interestingly, the variations and time-scale of anti-persistence are consistent with the variability of the strength of the wintertime atmospheric circulation over North Pacific (20-30 yrs). As such, the reported here anti-persistence constitutes a new and additional piece in the puzzle of interdecadal oscillations in the climate system.

Acknowledgments

The authors acknowledge NCEP/NCAR for reanalysis, NOAA/NESDIS for paleoclimate data and the following funding supports: L. M. V. Carvalho FAPESP (Proc. 02/09289-9), CNPq (302203/02-8, 474033/04-0); A. A. Tsonis ATM-0438612; C. Jones NOAA/CLIVAR NA050AR4311129.

References

Alley, R. B., J. Marotzke, W. D. Nordhaus, J. T. Overpeck, D. M. Peteet, R. A. Pielke Jr., R. T. Pierrehumbert, P. B. Rhines, T. F. Stocker, L. D. Talley, and J. M. Wallace (2003), Abrupt climate change. Science, 299, 2005-2010.

Baba, K., S. Minobe, N. Kimura, and M. Wakatsuchi (2006), Intraseasonal variability of sea-ice concentration in the Antarctic with particular emphasis on wind effect, J. Geophys. Res., 111, doi:10.1029/2005JC003052.

Beniston, M. (2004), The 2003 heat wave in Europe: A shape of things to come? An analysis based on Swiss climatological data and model simulations. Geophys. Res. Lett. 31, doi:10.1029/2003GL018857.

Briffa, K. R., P. D. Jones, and F. H. Schweingruber (1988), Summer Temperature Patterns over Europe - a Reconstruction from 1750-Ad Based on Maximum Latewood Density Indexes of Conifers. Quaternary Research, 30, 36-52.

Deser, C., A. S. Phillips, and J. W. Hurrell (2004), Pacific interdecadal climate variability, Linkages between the tropics and the North Pacific during boreal winter since 1900. J. Climate 17, 3109-3124

Emanuel, K. (2005), Increasing destructiveness of tropical cyclones over the past 30 years, Nature, doi: 10.1038/nature03906

Feder, J. (1988): Fractals, Plenum, New York.

Frankignoul, C., E. Kestenare, N. Sennechael, G. de Coetlogon, and F. D'Andrea (2000), On decadal-scale ocean-atmosphere interactions in the extended ECHAM1/LSG climate simulation. Climate Dyn., 16, 333-354.

Gu, D. F. and S. G. H. Philander (1997), Interdecadal climate fluctuations that depend on exchanges between the tropics and extratropics, Science, 275, 805-807.

Holton, J. (2004) An introduction to dynamic meteorology, fourth edition. 531 pp, Elsevier, Academic Press.

IPCC (2001): Climate change 2001: the scientific basis. Contribution of working group I to the third assessment report of the Intergovernmental Panel on Climate Change. Houghton, J. T., et al. (eds.), 881 pp, Cambridge University Press, Cambridge UK

Kalnay, E., M. Kanamitsu, R. Kistler,W. Collins,D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu , A. Leetmaa, B. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, Roy Jenne, and Dennis Joseph et al. (1996), The NCEP-NCAR 40 Year Reanalysis Project. Bull. Amer. Meteor. Soc., 77, 437-471.

Karl, T. R. , R. W. Knight, and N. Plummer (1995), Trends in High-Frequency Climate Variability in the 20th-Century. Nature, 377, 217.

Karl, T. R. and R. W. Knight (1997), The 1995 Chicago heat wave: how likely is a recurrence? Bull. Amer. Meteor. Soc., 78, 1107-1119.

Knutson, T. R., and S. Manabe (1998), Model assessment of decadal variability and trends in the tropical Pacific Ocean, J. Climate, 11, 2273-2296.

Latif, M. and T. P. Barnett (1996), Decadal climate variability over the North Pacific and North America: Dynamics and predictability. J. Climate, 9, 2407-2423.

Lau, W. K. M. and D. E. Waliser (2005), Intraseasonal Variability in the atmosphere-ocean climate system. 436 pp, Springer, NY

Madden, R. A. and P. R. Julian (1994), Detection of a 40-50 day oscillation in the tropics. J. Atmos. Sci., 43, 3138-3158.

Mann, M. E.,R. S. Bradley, M. K. Hughes (1998), Global-scale temperature patterns and climate forcing over the past six centuries. Nature, 392, 779.

McPhaden, M. J., D. X. Zhang (2002), Slowdown of the meridional overturning circulation in the upper Pacific Ocean, Nature, 415, 603-608.

Meehl, G. A. and C. Tebaldi (2004), More intense, more frequent, and longer lasting heat waves in the 21st century. Science, 305, 994-997.

Pierce, D. W., T. P. Barnett, N. Schneider, R. Saravanan, D. Dommenget, and M. Latif (2001), The role of ocean dynamics in producing decadal climate variability in the North Pacific., Climate Dyn., 18, 51-70.

Peixoto, J. P., and A. H. Oort, 1992: Physics of Climate. American Institute of Physics, New York, NY, 520 pp

Przybylak, R. (2002). Changes in seasonal and annual high-frequency air temperature variability in the Arctic from 1951 to 1990. Int. J. Climatol, 22, , 1017-1032

Serreze, M. C., J. A. Maslanik, T. A. Scambos, F. Fetterer, J. Stroeve, K. Knowles, C. Fowler, S. Drobot, R. G. Barry, and T. M. Haran (2003), A record minimum arctic sea ice extent and area in 2002, Geophys. Res. Lett., 30(3), 1110, doi:10.1029/2002GL016406.

Shepherd, A. ,D. Wingham, T. Payne, P. Skvarca (2003), Larsen ice shelf has progressively thinned. Science, 302, 856.

Trenberth, K. E., J. W. Hurrell (1994), Decadal Atmosphere-Ocean Variations in the Pacific. Climate. Dyn., 9, 303-319.

Tsonis, A. A. (Ed.) (1992), Chaos from theory to applications, 274 pp, Plenum, NY.

Tsonis AA, Roebber PJ, Elsner JB (1998) A characteristic time scale in the global temperature record. Geophys. Res. Lett. 25: 2821–2823

Tsonis, A. A., A. G. Hunt, and J. B. Elsner (2003), On the relation between ENSO and global climate change. Meteor. Atmos. Phys., 84, 229-242.

Xiong, L. , and J. G. Palmer (2000),Reconstruction of New Zealand temperatures back to AD 1720 using Libocedrus bidwillii tree-rings. Climatic Change, 45, 339.

22 Time Evolution of the Fractal Dimension of Electric Self-Potential Time Series

F. Cervantes-De la Torre [1, *], C. G. Pavía-Miller [2, 3], A. Ramirez-Rojas [2], F. Angulo-Brown[3]

[1] Universidad Autónoma Metropolitana Azcapotzalco, Departamento de Sistemas, Area de Sistemas Computacionales Av. San Pablo 180, Reynosa-Tamaulipas, 02200, México D.F. México. e-mail: fcdt@correo.azc.uam.mx

[2] Universidad Autónoma Metropolitana Azcapotzalco, Departamento de Ciencias Básicas, Area de Física de Procesos Irreversibles. Av. San Pablo 180, Reynosa-Tamaulipas, 02200, México D.F. México

[3] Depto de Física, Escuela Superior de Física y Matemáticas, Instituto Politécnico Nacional. Edif 9, U.P. Zacatenco, 07738, México D.F., México

Abstract. A statistical analysis of electric self-potential time series measured in an active seismic region of the state of Guerrero Mexico is presented. The fractal dimension D of these time series is calculated using Higuchi's algorithm. Most of the time the self-potential time series have $D \approx 2$, corresponding to uncorrelated white noise. However, there exist time intervals where D is remarkably less than two, corresponding to correlated signals. There are no clear correlations between D-changes and impending earthquakes. Nevertheless, some weeks before an earthquake $M = 7.4$ with epicenter 110 km from Acapulco, D reached its lowest value in almost four years. The possibility that D corresponds to charge patterns arising from self-organized critical phenomena is discussed.

1 Introduction

In seismically actives zones around the world there exist research programs for the study of precursory phenomena of seisms (Lomnitz 1990; Rikitake 1976; Hayakawa 1999). For over 25 years, one of the techniques used in the search of precursor phenomena of seisms consists in monitoring the so-called electric self-potential field. This field is studied through measurements of the ground electrical potential (the self-potential) by means of shallow pairs of unpolarized electrodes buried in the ground generating a voltage time series $\Delta V = \Delta V(t)$.

Several authors have proposed a correlation between patterns of self-potential variations and the mechanism of preparation of earthquakes (Varotsos and Alexoupoulos 1984a; 1984b). Some years ago, measures made near reservoir lakes in the French Alpes (Trique et al. 1999) have shown a correlation between the deformation of the earth crust and the variation of two physical parameters, namely, radon emission and electric potential variations. These authors have registered a fall of the electrical signal in one of the couples of electrodes. Although the recordings were made in a context different to that of a seismic zone, this correlation among a mechanical incitement of the earth crust and electrical signals is also relevant for the seismic

analysis. Since some years ago, we have taken registers of the fluctuations of the electric self-potential of the ground in several sites of Mexico (Yépez et al. 1995; Yépez et al. 1999), these registers were taken by means of electric self-potential stations as described in section 2. Some stations are located along the coast of Guerrero state, near the Middle American trench, which is the border between the Cocos and the American tectonic plates. Another stations (control stations) were located at Cholula Puebla (19° N, 98.3° W) and Mexico City (19.5° N, 99.1° W) more than 300 Km. distant from the trench. The first one is a very seismically active zone (Singh et al 1983), while the second one has a moderate seismicity (González and Valdés 1995).

In this work, we study the electric self-potential time series arising from the station of Acapulco (16.85° N, 99.9° W) linked to the Middle American trench. In this typical station, thousands of data are taken each two (or four) seconds during periods in the scale of months and years. These time series have been analyzed by means of several methods as the spectral ones (Yépez et al. 1995; 1999). Here we report on a study of the electric self-potential time series analyzed by means of the fractal dimension D computed by using the Higuchi´s algorithm (1988). The paper is organized as follows: In section 2 we present the advantages of the fractal dimension for characterizing the electric self-potential time series; in section 3 we show the temporal evolution of the fractal dimension; in section 4 we discuss about the limitations of our approach and finally we present some concluding remarks.

2 Fractal Dimension of Electrotelluric Time Series

The electric field derived from the electric potential difference between electrodes buried in the ground has been measured for a long time as one of the means of searching for anomalous phenomena possibly associated with impending earthquakes (Honkura 1981). This kind of electric features, frequently called self-potential anomalies [SPA] have been observed in laboratory experiments with typical crust's rocks (as granite) subject to increasing compressions (the SPA were present preceding and accompanying the sample destruction event (Sobolev et al. 1982). SPA observed in field measurements have also been reported possibly related with impending earthquakes [EQ] in China (Zhang and Fu 1981), Greece (Varotsos and Lazaridou 1991), U.S.A. (Corwin and Morrison 1977), Russia (Sobolev 1975) and Mexico (Yépez et al. 1995). In spite of these results, the SPA-method is considered as an EQ-prediction procedure susceptible of confusing interpretations because of the difficulties in separating SPA directly related to the mechanism of EQ preparation from those due to other sources such as noises of natural and artificial origins. However, in many seismically actives zones of the world, attempts are pursued for understanding this kind of possible EQ-precursor phenomena (Parrot et al. 1993, Hayakawa 1999).

The electric self-potential field is usually measured by means of shallow pairs of unpolarized electrodes buried in the ground. The distance between electrodes is typically between 50 and 2500 m (Varotsos and Lazaridou 1991). In our case, we use stainless steel electrodes buried 2 m into the ground and 50 m apart. We registered

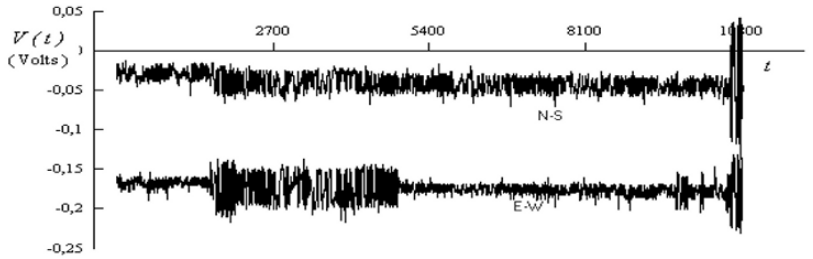


Fig. 22.1. Typical segments of Self-Potential time series collected at Acapulco-station. The superior series is the N-S one and the down series corresponds to the E-W line. Here, two 6 hr series are showed, with a sampling period of 2s

the voltage between the electrodes by a filter and amplifier coupled to an analog to digital converter (PCL 711, 12 bits), the data were stored in a magnetic disk every 4 sec or 2 sec. The frequency cutoff of the filter was set at 5 Hz the filters were designed in such a way that their response was flat up to 5 Hz and falling to a negligible value at 10 Hz; in this way we assure the elimination of the influence of the power line frequency (60 Hz). With the above instrumental array we collected ΔV-data during time intervals in the scale of months and years.

Typical segments of our $\Delta V(t)$ time series is depicted in Fig. 22.1. The behavior of these time series did not show any dependence with rainfall (Yépez et al., 1995). Since 1992, we have obtained $\Delta V(t)$ time series at several sites (electric self-potential stations) during diverse periods, along the Pacific coast of Guerrero state in southern Mexico, a very active seismic zone linked to the Middle American Trench. During this period at the vicinity of our stations ($r \leq 200$ Km for each station) have occurred 10 EQ's with $M_w \geq 5.9$ (Rudolf-Navarro, 1999). The strongest of these EQ's ($M_w = 7.4$) occurred in September 14, 1995 with epicenter at (16.31° N, 98.88° W), 110 Km distant from our station at Acapulco. Because this has been the most relevant event near our stations, we devote the fractal analysis to it.

The Acapulco-station was installed in October 16, 1992. Since that date, it has continuously worked with only a few interruptions due to hurricanes and technical causes. These data sets constitute the so-called electric self-potential time series $\Delta V(t)$. The theory of nonlinear dynamical systems provides new tools and quantities for the characterization of irregular time series data, as the electric self-potential time series. In this paper, we show that measurings of the fractal dimension D by means of the Higuchi´s method are appropriated for characterizing the time series. We study the temporal evolution of the fractal dimension D and we discuss the feasibility of this method to identify patterns that would be precursors of seisms with relevant magnitudes. Many time series obtained from discrete measurements on physical systems and in observations in the field display statistically self-affine properties.

Self-affine fractals are generally treated quantitatively using spectral techniques, for a time series that is fractal, the power spectral density S(f) has a power law dependence on frequency f (Turcotte 1992), that is

$$S(f) \propto f^{-\alpha}, \tag{1}$$

where α is an exponent related with the fractal dimension D of the time series, and with the exponent H known as the Hausdorff measure, by means of (Turcotte 1992; Burlaga and Klein 1986)

$$\alpha = 2H + 1 = 5 - 2D. \tag{2}$$

In our study of electric self-potential time series, we use the Higuchi´s algorithm for calculating their fractal dimension D. Usually, the exponent α is considered to be the index for representing the irregularity of a time series, although the fractal dimension D can be also used as index of irregularity. In fact, in many cases the usage of D is more appropriate than the spectral exponent α for determining irregularity indices (Cervantes et al. 1999). The fractal technique developed by Higuchi (1988) gives stable indices even for a small number of data. Higuchi (1988; 1990) considers a finite set of time series observations taken at a regular interval: $\Delta V(1), \Delta V(2), \Delta V(3), \cdots\cdots, \Delta V(n)$. From the given time series, he first constructs a new time series, ΔV_k^m, defined as follows

$$\Delta V_k^m : \quad \Delta V(m), \Delta V(m+k), \Delta V(m+2k), \ldots, \Delta V\left(m + \left[\tfrac{N-k}{k}\right].k\right), \tag{3}$$

with m = 1, 2, …, k, and where [ ] denotes the Gauss notation, and m, k are integers that indicate the initial time and the time interval respectively. For a time interval equal to k, one gets k sets of new time series. Higuchi defines (1988) the length of the curve associated to each time series ΔV_k^m as follows:

$$L_m(k) = \frac{\sum_{i=1}^{\left[\frac{N-m}{k}\right]} \left[\Delta V(m+ik) - \Delta V(m+(i-1)k)\right] \left(\frac{N-1}{\left[\frac{N-m}{k}\right]k}\right)}{k} \tag{4}$$

where the term $\frac{(N-1)}{\left[\frac{N-m}{k}\right]k}$ represents a normalization factor, then the length of the curve for the time interval k is taken as the average value $\langle L(k) \rangle$ over k sets of $L_m(k)$. If the average value obeys the following scaling law:

$$\langle L(k) \rangle \propto k^{-D}, \tag{5}$$

then the curve is fractal with dimension D (Higuchi 1988). This algorithm can be applied even over time series that are not stationary and this fact represents an advantage over the spectral techniques (Cervantes et al. 1999). Higuchi (1990) shows that if $1 \le \alpha \le 3$ then $D = \frac{5-\alpha}{2}$ (equation (2)) is held, he also shows that the following limits are held,

$$if \quad \alpha \to 0 \quad then \quad D \to 2\,, \tag{6}$$

which corresponds to uncorrelated white noise, while the second limit is: if $\alpha \to 3$ then $D \to 1$.

Since the values of the fractal dimension D for our time series are in the interval [1,2], the equations (2) and (6) can be used for their analysis. For the calculation of the fractal dimension D, by means of the Higuchi algorithm we divide the $\Delta V(t)$ time series data in 6 hrs-files (10800 ΔV-points) (if the division is smaller, the results do not change appreciably). To each file, we associate a fractal dimension D taken from (5), calculating the slope of the double log plot of $\langle L(k) \rangle$ against k. Applying the Higuchi´s algorithm to data as those of Fig. 22.1, we can see in Figs. 22.2a and 22.2b, that we obtain straight lines which have a very accurate slope (D) when are calculated by means of a least square fit, in contrast with typical cases of log-log plots of power spectra versus frequency (Cervantes et al. 1999).

It is convenient to remark that in the studied series the calculation of the fractal dimension D is statistically appropriate; that is, the correlation coefficient R^2 is near to one, and the associate errors of the fractal dimension D are random. That is, the

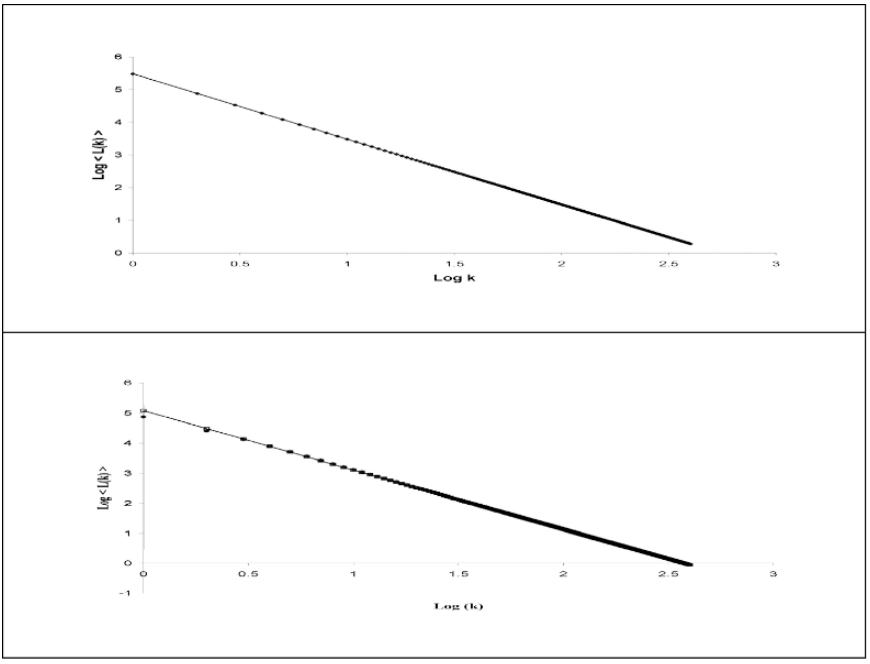


Fig. 22.2. a).- Log-log plot of the average length against k for a 6-hr file (E-W line) corresponding to October 19, 1993. The fractal dimension is the slope of the straight line with a value D = 1.99906 ± 0.00034 and R^2 = 0.99999. This series corresponds to a typical white noise period. b).- The same of Fig. 22.2a, for a segmente (April 5, 1997) of correlated noise with D = 1.9745774 ± 0.0015 and R^2 = 0.999768.

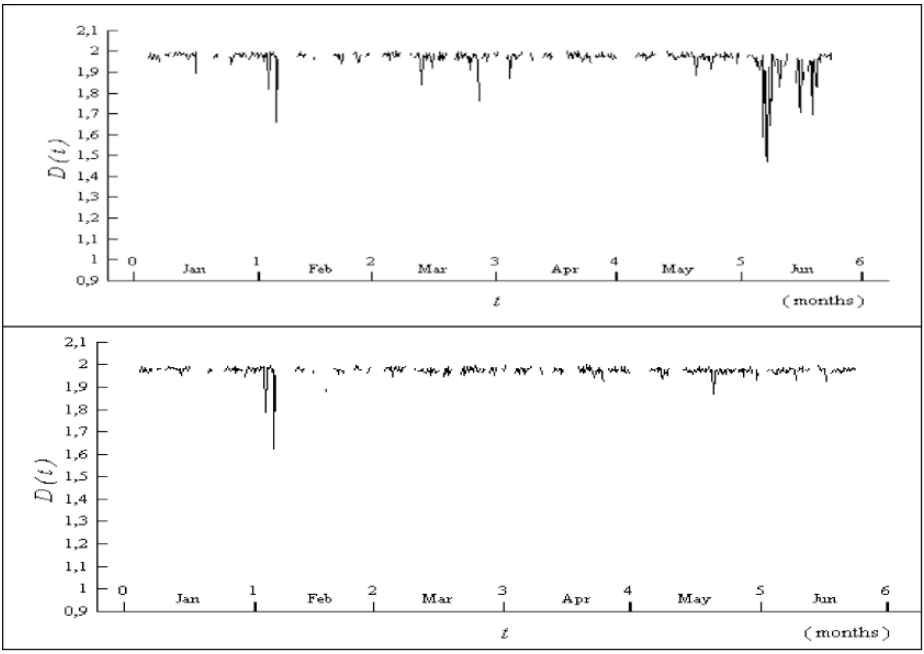


Fig. 22.3. a).- Time evolution of fractal dimension *D* for the E-W short line from January to June of 1998. During this last month the short line presents remarkable fluctuations that are not present in the E-W long line depicted below. b) The fluctuations in 3a may correspond to very local noises in one of the electrodes of the E-W short line. Each point of the plot corresponds to a 6 hr-file of the self-potential time series.

obtained graphs of $\log \langle L(k) \rangle$ vs. $\log k$ for our electric self-potential time series are always statistically straight lines. Since January 1998, we installed at Acapulco station two short lines (N-S, E-W) with a separation of 25 m between electrodes. The original old lines are 50 m apart. In Figs. 22.3a and 22.3b, we see the time evolution of the fractal dimension *D* for both E-W parallel lines during a six months period. We observe that the electric self-potential time series taken from Acapulco station stay most of the time with $D \approx 2$, that is, the time series mainly correspond to uncorrelated white noise.

Nevertheless, there exists a certain number of time intervals, where *D* is remarkably less than two, corresponding to the range where α and *D* are linked by eq. (2), that is, corresponding to a correlated signal. It is remarkable that at the end of June, 1998, the short line shows some *D*-fluctuations that are not present in the 50 m line. This same behavior but with less intensity is observed in the N-S parallel lines (Figs. 22.4a and 22.4b). We believe that these differences between long and short parallel lines would help to distinguish local perturbations over the electrodes, such as the electrochemical ones (Varotsos and Lazaridou 1991). We have shown that the usage of the Higuchi´s method leads to very precise values of the fractal dimension

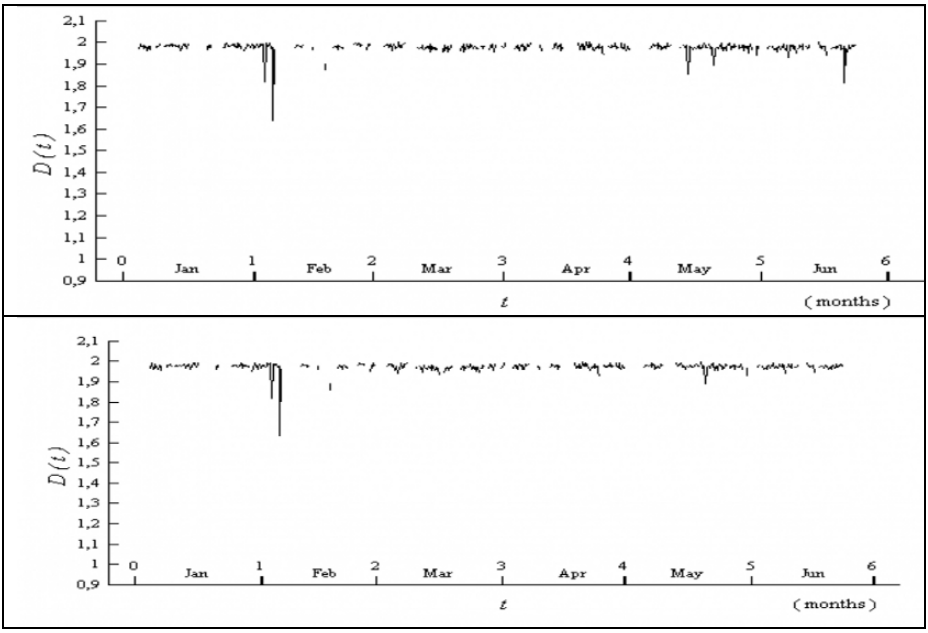


Fig. 22.4. a).- Time evolution of D for the N-S short line; b).- Time evolution of D for the N-S long line. In this case the behavior of both lines is very similar. Each point of the plot corresponds to a 6 hr-file of the self-potential time series.

D of the electric self-potential time series (Cervantes et al. 1999). Therefore, the electric self-potential time series obey the universal scaling law (5), by means of which we can determine the fractal dimension D, which describes some properties of the dynamic structure of the electric time series.

3 Time evolution of the Fractal Dimension

The theory of nonlinear dynamical systems provides new tools and quantities for the characterization of irregular time series data through universal scaling laws, as the scaling law proposed by Higuchi (5). These tools yield an approximate description of the structure of a system by means of statistical techniques which are useful for classification and diagnosis, even in situations where determinism is almost lacking. In this work, we present some studies of electric self-potentials arising from the station of Acapulco, we characterize the voltage time series by means of the fractal dimension D.

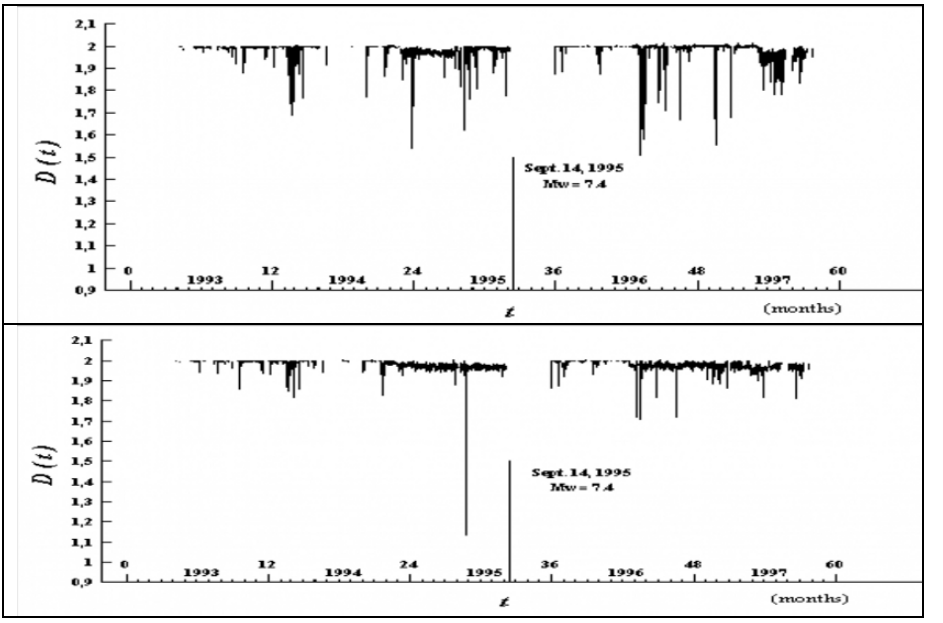


Fig. 22.5. a).- Time evolution graph of fractal dimension *D* for a period from March, 1993 to December, 1997 (E-W line). Some few intervals lacking of data are depicted. Most of the time the value of *D* is close to two (white noise) and there exist time intervals where *D* is remarkably less than two (correlated signals). b).- The same of Fig. 22.5a for the N-S line. Some weeks before the Mw = 7.4 earthquake (marked with an arrow) *D* reached its lower value *D* = 1.136. Each point of the plot corresponds to a 6 hr-file of the self-potential time series.

For the Acapulco station, we study the behavior of the fractal dimension and we discuss the feasibility of the method to identify patterns that would be precursors of seisms. In Figs. 22.5a and 22.5b we show the time evolution of the fractal dimension *D* since 1993 up to 1997, in such graphs we observe that electric self-potential time series stay most of the time with $D \approx 2$, which describes uncorrelated white noise, the *D*-value for other time intervals are smaller than two but bigger than one, after these short intervals, the dominant effect of the white noise appears again. These changes in the fractal dimension *D* are associated with changes in the dynamics associated to the electric self-potential time series. The intervals with $D \neq 2$ mainly correspond to *D* values between 1.5 and 1.85 approximately, corresponding to values of the spectral exponent α between 1.3 and 2 (see eq. 2); that is, values in the range of fractional Brownian motion (Turcotte 1992). Both intervals for D and α are within the values recently reported by Hayakawa et al. (1999) in the case of ULF geomagnetic data associated to the Guam EQ on August 8, 1993. If α is in the interval [1.3,2], then the *H* exponent is in the interval [0.15,0.5] (see eq. 2), indicating that the process has long range anticorrelations or antipersistence (Ranzarajan and Ding 2000). That is, the signal has a behavior such that an increasing trend in the past implies a decreasing trend in the future and vice versa. For Brownian noise, $H = \frac{1}{2}$, $D = \frac{3}{2}$ and $\alpha = 2$.

Since we attempt to correlate the occurrence of one seism with the behavior of the fractal dimension, we are interested in investigating those intervals of time where the fractal dimension *D* changes drastically (Hayakawa et al 1999). During the studied period at the vicinity of our stations the strongest EQ, $M_w = 7.4$ occurred in September 14, 1995 with epicenter at (16.31 °N, 98.88°W), 110 Km distant from our station at Acapulco (16.9 °N, 99.9 °W). Because of this has been the most relevant event nearly our stations, we shall mainly devote to the fractal analysis of it. It is remarkable that the most conspicuous change of *D* appeared linked to the main earthquake in the monitored zone during four years (see Fig. 22.5b). In fact, the smallest *D*-value obtained in the studied period was $D = 1.136$ in the N-S line around May 26 , 1995, that is, three and a half months before the EQ, $M_w = 7.4$. This change was not observed in the E-W line. This kind of asymmetry between the E-W and N-S lines has been previously reported (Varotsos and Lazaridou 1991; Yépez et al. 1995; Hayakawa et al. 1999). We have not a plausible explanation of this fact, but we believe that it could be very interesting if the fractal dimension *D* of electric self-potential time series were measured in some other regions of the world where electric self-potential stations are at working.

4 Discussion

Since October 1992, a network of three electrotelluric stations (Acapulco, Coyuca and Ometepec I) were first installed (Yépez et al. 1995) along the coast of Guerrero State in Southern Mexico. Later, in 1994 other three stations were added to that network (San Marcos, Tecpan and Ometepec II). Since 1992 and 1994 due to different technical and logistic problems, the station which has worked with more regularity is the Acapulco station. For these reasons, in this paper we only report the fractal analysis of that station. However, in Cervantes (1998) a brief fractal analysis for short periods for Coyuca and Ometepec I stations are reported. Unfortunately, these periods do not include the September 1995 earthquakes data.

The Guerrero State coast is a very active seismic zone where seisms of all sizes occur. However, the irregularity of the self-potential time series is so big, that is a very difficult task to identify small variations in connection with small or medium seisms. We have adopted the same strategy that other authors (Hayakawa et al. 1999) consisting in the search of possible connections between electrotelluric anomalies and EQ's only in the case of great events. For the M=7.4 EQ occurred in September 14, 1995 at 110 Km from Acapulco station, we can not assert that the fractal anomaly observed in Fig. 22.5b ($D \approx 1.136$) is a precursory signal of that event. However, we want to emphasize that a fractal analysis of more electrotelluric data measured in several seismic zones around the world is necessary for the search of insights linked to the usage of fractal methods. In Fig. 22.3, each point of the plot corresponds to 6 hr.-files of electrotelluric data. Thus, the D=1.136 point has this duration. We refer to this value as an anomalous one based only in the fact of its singularity, that is, in almost four years it is a very inusual value. Most of the time (before and after the main EQ) we have a fractal dimension $D \approx 2$ (uncorrelated white noise) with fluctuations in the intervals (1.7,2) for the NS-line and (1.5,2) for the EW-line respectively. Thus, D=1.136 can be considered as an anomalous value. Until now, we have not

clear arguments for interpreting this anomaly as a true precursory phenomenon to the 7.4 EQ, however, in the next section we make a conjecture about this matter in the context of self organized critical processes.

5 Concluding Remarks

In general, the data in log-log plots of $\langle L(k) \rangle$ versus k obtained from our measurements are aligned along straight lines and a robust power-law emerges, Eq. (5). In this paper, we study the dynamical behavior of the fractal dimension D, and we have found that apparently the value $D \approx 1.136$ is possibly linked with the occurrence of a seism with magnitude $M_w = 7.4$. For larger time periods the typical behavior of the electric-self potential is of the white noise-type with D = 2. Nevertheless, there exist some periods where D is notoriously less than 2 corresponding to fractional Brownian noise. In 1987, Bak et al. proposed the notion of self-organized criticality as a general organizing principle governing the behavior of spatially extended dynamical systems with both temporal and spatial degrees of freedom. According to this principle, composite open systems having many interacting elements organize themselves into a stationary critical state with no length or time scales others than those imposed by the finite size of the system. The critical state is characterized by spatial and temporal power laws.

According to Bak et al. (1987, 1988, 1989), the temporal "fingerprint" of the SOC-state is the presence of $\frac{1}{f^{\alpha}}$ noise and its spatial signature is the emergence of scale invariant (fractal) structure. Since the Bak et al. (1987, 1988), pioneering papers, numerous SOC physical realizations have been proposed (Machta et al. 1993; Vattay and Harnos 1994; Marani et al. 1998; Bezryadin et al. 1999). For instance, Sornette and Sornette (1989) suggested that SOC is relevant for understanding EQ as a relaxation mechanism which organizes the crust both at spatial and temporal levels. This idea allows us to rationalize observations on occurrences and magnitudes of EQ. Self-organized criticality represents an attempt towards the elucidation of a general mechanism responsible for the ubiquitous appearance of natural phenomena lacking a spatial or temporal scale (Marani et al. 1998). Between the great variety of phenomena showing scale and time invariance are several geophysical processes (Turcotte 1992; Angulo-Brown et al. 1998). In particular, self-potential time series show power laws with values of D, α and H, in some cases within the ranges of typical SOC-phenomena. In 1998, Marani et al. proposed a model for spontaneous aggregation of charged metal balls floating in viscous liquid with a high dielectric constant. In the Marani et al. -model the structure having minimum total potential energy is characterized by a great regularity. This minimum energy is substantially lower than the values obtained by spontaneous self-organization, and the charge distribution correspond to a fractal dimension $D = 2$. On the other hand, these authors find dendritic structures of charges that represent the emergency of fractality ($D < 2$) as a result of the existence of a set of stationary or recursive states derived from the dissipation of the injected charge through the boundaries of the system which impose a global constraint which results in the metastable states becoming scale free over a certain range of the length scale. In summary, the aggregates of the

Marani et al.-model change their structures between $D = 2$ and D´s as low as $D = 1.3$. These results are reminiscent of the behavior of our self-potential time series, suggesting this likely, that within the underground there exist a charge distribution that changes between a base state ($D = 2$) corresponding to a uniform distribution and non uniform distributions ($D < 2$) with bigger potential energy. Some authors (Sornette and Sornette 1990) have conceived the crustal medium as a kind of complex ceramic with dielectric and hydrostatic behavior. We want to emphasize that this assertion is only a conjecture and more studies and analysis are necessary for understanding the behavior and meaning of the fractal dimension of self-potential time series.

Acknowledgments

* Student of Universidad Autonoma del Estado de México, Doctorado en Física No-Lineal, Facultad de Ciencias.

References

Angulo-Brown, F., A. H. Ramirez-Guzmán, E. Yépez, A. Rudolf-Navarro y C. G. Pavía-Miller, Fractal geometry and seismicity in the Mexican subduction zone, *Geofis. Int*., **37**: 29-33, 1998.

Bak P., C. Tang and K. Wiesenfeld, Self-organized criticality: an explanation of 1/f noise, *Phys. Rev Lett*., **59**, 381-384, 1987; Self-organized criticality, *Phys Rev. A.,* **38**, 364-374, 1988.

Bak P. and C. Tang, Earthquake as a self-organized critical phenomenon, *J. Geophys. Res.,* **54**, 15635-15637, 1989.

Bezryadin A., M. R. Westervelt and M. Tinkham, Evolution of avalanche conducting states in electrorheological liquids, *Phys. Rev. E*, **59**, 6896-6902, 1999.

Burlaga L.F. and L.W. Klein, Fractal structure of the interplanetary magnetic field, *J. Geophys. Res.,* **91A1**, 347-350, 1986.

Cervantes de la Torre F., A. Ramirez-Rojas, C. G. Pavía-Miller, F. Angulo-Brown, E. Yépez and J. A. Peralta, A comparison between spectral and fractal methods in electrotelluric time series, *Rev. Mex Fis.*, **45**, 298-302, 1999.

Cervantes De la Torre Francisco. Análisis fractal de series de tiempo electrotelúricas., *M. Sci. Thesis, ESFM-IPN*, 1998.

Corwin R.F. and H.F. Morrison, Self potential variation preceding earthquakes in central California, *Geophys. Res. Lett.,* **4**, 171-174, 1977.

González-Pomposo G. J. y Valdés-González C., Seismicity of the State of Puebla Mexico 1986-1989. *Geofis. Int*., **34**: 221-231, 1995.

Hayakawa Masashi, Ito Tetsuya and Smirnova. Natalia. Fractal analysis of ULF geomagnetic data associated with the Guam earthquake on August, 8, 199. Geophysical Research Letters, Vol. 26, No. 18. 2797-2800, September 1999.

Hayakawa M., Editor., *Atmospheric and Ionospheric Electromagnetic Phenomena Associated with Earthquakes*., Terra Sci. Pub., Tokyo. 1999.

Higuchi T., Approach to an irregular time series on basis of the fractal theory, *Physica D.,* **31,** 277-283, 1988.

Higuchi T., Relationship between the fractal dimension and the power law index for a time series: a numerical investigation, *Physica D.,* **46**, 254-264, 1990.

Honkura Y., Electric and magnetic approach to earthquake prediction, *in Earthquake Prediction I,* pp. 301-382, Editor. T. Rikitake, D. Reidel Publ. Co, Boston, 1981.

Lomnitz C., *Fundamentals of Earthquake Prediction*, John Wiley & Sons., 1990.

Machta J., D. Candela and R.B Hallock, Self-organized criticality in ^{4}He with a heat current, *Phys. Rev. E.,* **47**, 4581-4581, 1993.

Marani M., J. R. Banavar G. Caldarelli, A Maritan and A. Rinaldo, Stationary Self-organized fractal structures in an open dissipative electrical system*, J. Phys. A: Math. Gen.,* **31**, l337-l343, 1998.

Parrot M., J. Archache, T.J. Berthelier, E. Blanc, A. Deschamps, F. Lefeuvre, M. Menvielle, J.L. Plantet, P. Tarits and J.P. Villain, High-frequency seismo-electromagnetic effects, *Phys. Earth Planet. Int.*, **77**, 65-83, 1993.

Rangarajan Govindan and Ding Mingzhou. Integrated approach to assessment of long range correlation in time series data. Physical Review E. Vol. 61. No. 5. May 2000.

Rikitake T., *Earthquake Prediction* Developments in Solid Earth Geophysics 9, Elsevier Scientific Publishing Company., 1976.

Rudolf-Navarro A. H., Patrones de quietud sísmica como precursores de grandes sismos en México, *M. Sci. Thesis, ESFM-IPN*, 1999.

Singh S. K., M. Rodriguez and L. Esteva., Statistics of small earthquakes and frequency of occurrence of large earthquakes along the mexican subduction zone., *Bull. Seism. Soc. Am.*, **74**: 267-279, 1983.

Sobolev G.A., Application of electric method to the tentative short-term forecast of Kamchatka earthquakes, *Pure Appl Geophys.*, **113**, 229-235, 1975.

Sobolev G.A. et al., Precursors of the destruction of large rock samples, *Izv. Earth Phys.*, **18**, 572-580, 1982.

Sornette A. and D. Sornette, Self-organized criticality and earthquakes, *Europhys. Lett.*, **9**, 197-202, 1989.

Sornette A. and D. Sornette, Earthquake rupture as a critical point: consequences for telluric precursors, *Tectonophysics.*, **179**, 327-334, 1990.

Trique M., P. Richon, F. Perrier, J. P. Avouac and J. C. Sabroux, Radon emanation and electric potential variations associated with transient deformation near reservoir lakes., *Nature.*,**399**, 137-141, 1999.

Turcotte D.L., *Fractals and chaos in geology and geophysics*, Cambrige University Press, 1992.

Varotsos P. and K. Alexoupoulos, Physical Properties of the Variations of the electric field of the Earth preceding earthquakes I., *Tectonophysics.*, **110**, 73-98, 1984a.

Varotsos P. and K. Alexoupoulos, Physical Properties of the Variations of the eletric field of the Earth preceding earthquakes II., *Tectonophysics.*, **110**, 99-125, 1984b.

Varotsos P. and M. Lazaridou, Latest aspects of earthquake prediction in Greece based on seismic electric signals, *Tectonophysics.*, **188**, 321-347, 1991.

Vattay G. and A. Harnos, Scaling behavior in daily air humidityt fluctuantions, *Phys. Rev. Lett.*, **73**, 768-771, 1994.

Yépez E., F. Angulo-Brown, J.A. Peralta, C.G. Pavía-Miller and G.González-Santos, Electric field patterns as seismic precursors, *Geophys. Res. Lett.*, **22**, 3087-3090, 1995.

Yépez E., J. G. Pineda, J. A. Peralta, A. V. Porta, C. G. Pavía-Miller and F. Angulo-Brown, Spectral Análisis of ULF Electric Signals Possibly Associated to Earthquakes., *Atmosferic and Ionospheric Electromagnetic Phenomena Associated with Earthquakes*. Ed. M. Hayakawa, Terra Sci. Pub., Tokyo., 115-121, 1999.

Zhang G. and Z. Fu, Some features of medium and short-term anomalies before great earthquakes, *in Earthquake Prediction*, pp. 497-509, Editors. Simpson, D.W. & Richards, P.G., American Geophysical Union, Washinton, DC, 1981.

23 Diffusion Entropy Analysis in Seismicity

Abigail Jiménez[1], Kristy F. Tiampo[2] and Antonio M. Posadas[3]

[1] University of Western Ontario, London, Canada, Department of Earth Sciences Biological and Geological Sciences, ajlloret@ual.es
[2] University of Western Ontario, London, Canada, Department of Earth Sciences Biological and Geological Sciences, ktiampo@seis.es.uwo.ca
[3] University of Almeria, Spain, Department of Applied Physics, aposadas@ual.es

Abstract. Long-memory dependence is associated with a Hurst coefficient, H, greater than a value of 1/2. Different techniques, previously tested with simulated time series, are used to calculate this coefficient for some seismic catalogues around the world, so that the results are reliable. However, these techniques assume a finite variance and according to the central limit theorem the underlying statistics are Gaussian. A complementary scaling analysis, the diffusion entropy analysis (DEA), that focuses on the scaling exponent δ evaluated through the Shannon entropy S(t) of the diffusion generated by the fluctuations of a time series, is performed in order to determine the characteristics of these time series.

1 Introduction

In the 1950s, Hurst developed a set of characteristic parameters in the study of Nile river flows and reservoir modelling (Hurst 1951; Hurst 1956). The ideal reservoir capacity $R(\tau)$ was divided by the standard deviation $S(\tau)$ of τ successive discharges. He noted that:

Except perhaps for small values of τ, the rescaled range $R(\tau)/S(\tau)$ is proportional to τ^H, where H is a constant between 0.5 and 1.

This stimulated (Mandelbrot and Wallis 1968) to introduce the concept of fractional Brownian motion (FBM). Problems related to the interpolation or prediction of processes exhibiting the Hurst phenomenon yield the following interpretations of H: A value H=1 corresponds to 1/f-noise or pink noise. The adoption of a colour name "pink" derives from the fact that a light source characterized by a 1/f spectrum looks pink. These type of noises are particularly important because they represent a balance between randomness and order, or between unpredictability and predictability. In fact, for pink noise, the autocorrelation function between two events separated by a time interval r is independent of r. Pink noises, H~1, are found in countless natural

phenomena from heartbeat intervals to music. A value $0<H<0.5$ corresponds to anti-persistent noise, $H=0.5$ corresponds to uncorrelated or random noise, also known as *white* noise, and $0.5<H<1$ corresponds to correlated or persistent noise. It is possible to extend the definition of H for values larger than 1. So, a value $H=1.5$ corresponds to Brownian motion, which, as it is well known, describes the erratic motion of a particle, such as a pollen grain, in suspension on a fluid; this erratic motion is caused by random collisions between the particle and the molecules of the fluid. A value $H=2$ corresponds to *brown* noise and a value $H>2$ is known as *black* noise. These noises are characterized by a very smooth shape and may be adopted, for example, to generate artificial mountain landscapes (Scafetta and West 2005).

Jiménez et al. (2006) used different techniques, tested with simulated time series, to calculate H in seismic catalogues: the Rescaled Range analysis (R/S), a method based on the discrete second-order derivative (DSOD), which can be also adapted to a wavelet framework with similar properties (WDSOD), and using the slope of the log-log plot of the detail variance vs. the level. These techniques assume a finite variance and according to the central limit theorem the underlying statistics are Gaussian.

Recently, Scafetta et al. (2001) introduced a complementary scaling analysis, the diffusion entropy analysis (DEA), that focuses on the scaling exponent δ evaluated through the Shannon entropy S(t) of the diffusion generated by the fluctuations of a time series. The diffusion generated by a fractal Gaussian intermittent noise is a particular type of FBM and satisfies the asymptotic scaling relation between indexes:

$$H = \delta \tag{1}$$

The presence of a Lévy-walk process in a given time series can be detected by means of the following asymptotic relation:

$$0.5 < \delta = (3-2H)^{-1} < 1 \tag{2}$$

So, by comparing H and δ, obtained by means of different techniques, we can decide if the time series follows a fractional Gaussian or Lévy behaviour. Scafetta and West (2004) suggested that earthquake clusters are 1/f long-range correlated, and that this implies that earthquake occurrences may strongly depend on the geological history of a vast region. In the present paper, we apply the DEA and the Hurst calculations to three seismic catalogues around the world, to evaluate the characteristics of these time series.

2 Methods

Hurst (1951) developed the Rescaled Range (R/S) analysis for estimating this dependence. The range R is the difference between the minimum and maximum accumulated values or cumulative sum of X (t, T) of the natural phenomenon at discrete integer-valued time t over a time span T, and the standard deviation S is estimated from the observed values $X_i(t)$. Hurst found that the ratio R/S is very well described for a large number of natural phenomena by the following empirical relation:

Table 23.1. Tests for different estimation techniques with different values of H H_{th} is the theoretical value of the simulated FBM). The DSOD, WDSOD and Wavelet methods found the same values.

H_{th}	DSOD-WDSOD-Wavelet	DEA
0.1	0.173 ± 0.021	0.145 ± 0.007
0.3	0.318 ± 0.021	0.303 ± 0.012
0.5	0.494 ± 0.018	0.492 ± 0.015
0.7	0.696 ± 0.016	0.682 ± 0.014
0.9	0.896 ± 0.015	0.847 ± 0016
1.0	0.997 ± 0.014	0.905 ± 0.013

$$R/S(t) = (ct)^H \tag{3}$$

where c is a constant. The R/S algorithm can be found, for example, in (Gimeno 2000). Some reasons for such a behaviour could be non-stationarity, pre-asymptotic behaviour or long-time correlations in stationary X(t, T) (Goltz 1997).

There are other methods for calculating the Hurst exponent, where the properties of the fractional Brownian motion (FBM) are used. The FBM is the mathematical model for a stationary process which exhibits the Hurst phenomenon. For example, the Aggregated Variance (AV) method is based on the main property of the self-similarity, i.e. the equality in distribution of two scaled samples at different time scales (Piron 2005); Istas and Lang (1994) use a method based on the discrete second-order derivative (DSOD), which can be also adapted to a wavelet framework with similar properties (WDSOD); finally, Abry et al. (2003) estimate H using the slope of the log-log plot of the detail variance versus the level. The implementation of the three last methods can be found in MATLAB, as well as utilities for simulating the FBM MATLAB. In Jiménez et al. (2006), these techniques are compared.

Here we test the DEA implemented following Scafetta et al. (2001). For testing the accuracy of these methods several FBM with different known H have been simulated. Because of the randomness of the simulations, 50 trials have been made for each one. In Table 23.1 we summarize the results, where the mean as well as the standard deviation for the 50 trials are given. As can be seen, the DEA is closer to the actual value for $H \leq 0.5$, approximately. However, for high values of H, the DSOD, WDSOD and Wavelet methods are fitter.

3 Data

We use three catalogues from the following regions: the Iberian Peninsula, Greece, and southern California, described below.

The Iberian Peninsula catalogue: The analyzed area is the region between 36º and 45º north latitude and between 10º west and 5º east longitude. The catalogue was obtained from the National Geographic Institute (Spain) and it contains all the seismic data in the Iberian Peninsula. Around 10000 earthquakes have been selected in

the period 1970-2001. The depths of the earthquakes range from 0 to 140 km, and the magnitudes are between 2.0 and 6.5 (mb and mbLg). The average errors in the hypocentral X, Y, and Z locations, are ±5 km, ±5 km and ±10 km, respectively, for the data recorded until 1985, and ±1 km, ±1 km and ±2 km for those acquired since 1985.

The Greek catalogue: Burton et al (2004a and 2004b) carefully created a homogeneous earthquake catalogue for Greece, because of the high seismicity in that region. The catalogue contains 5198 earthquakes during 1900-1999, which are within the area 33º-43º N, 18º-30.99º E, focal depths 1 to 215 km and magnitude range $4.0 \leq Mw \leq 7.7$. These authors used the catalogues and bulletins of the International Seismological Centre (ISC), and the research catalogues of Papazachos and Comninakis (1971), Makropoulos and Burton (1985), Makropoulos et al. (1989), Papazachos et al. (1994) and Papazachos and Papazachou (1997). The catalogue is complete above magnitude 5.0. All data have been used in the calculations, without removing foreshocks or aftershocks.

The Southern California catalogue: The catalogue is compiled by the Southern California Earthquake Center (SCEC) and contains the seismic data for the period 1932-1999. The analyzed area ranges from 32º to 39º north latitude, and 115º to 122º west longitude. The magnitude spans from 3.0 to 8.0.

4 Results

We can calculate the Hurst's exponent for the energy released and for the accumulated quantities derived from the magnitude of the events in a large area. From the relationship between magnitude and energy Gutenberg and Richter (1956), the following quantities can be analyzed:

$$\varepsilon_q(N(\tau)) = \sum_{n=1}^{N(\tau)} \varepsilon_n^q \tag{4}$$

where ε_n is the released energy of the n^{th} event, and $N(\tau)$ is the number of earthquakes in a given interval of time τ. Note that when $q = 1$, ε_q represents the accumulated energy; when $q = 1/2$, it is the Benioff 'strain' (Sammis and Smith 1999); for $q = 1/3$, it is an approximation of the fault slip dimension (Mai and Beroza 2000; Carpinteri and Pugno 2005); and for $q = 0$ it is simply the number of the events. The main proxy measures for strain are inferred from the radiated seismic energy, the seismic moment or the seismic event rate. The Benioff strain is based on essentially dimensional arguments, and is not directly related to a measurable tectonic strain, so that the physical meaning of the cumulative Benioff strain is not clear (Main 1999). However, it has been used in seismic hazard assessment (Bowman and Sammis 2004), and it is interesting to analyze its long-term behaviour. For the Hurst estimators described, the input time series is the integral over time of $\varepsilon^q{}_n$ (or cumulative $\varepsilon^q{}_n$ up to each time event in a discrete form), so that no τ is used, and no a priori smoothing

Table 23.2. Results for the Iberian Peninsula. q is the exponent in Eq. (4), R/S is the estimation for the Hurst exponent by means of the Rescaled Range method. The number in brackets is the correlation coefficient of the log-log linear fit. DSOD stands for the discrete second-order derivative method for obtaining the Hurst exponent as well. WDSOD is for the wavelet version of the last one. The fifth column is another wavelet estimation, and DEA is the value of δ obtained by means of the diffusion entropy analysis, with its corresponding correlation coefficient in the log-log plot.

q	R/S	DSOD	WDSOD	Wavelet	DEA
1	0.5116 ± 0.0013 (0.9997)	0.5383	0.5526	2.1231	0.035 ± 0.003 (0.8946)
1/2	0.633 ± 0.005 (0.9969)	0.6675	0.6799	1.2979	0.30 ± 0.03 (0.8921)
1/3	0.788 ± 0.006 (0.9973)	0.7042	0.7001	1.1279	0.87 ± 0.03 (0.9834)
0	0.956 ± 0.005 (0.9989)	0.0360	-0.0003	1.000	0.884 ± 0.006 (0.9991)
rate	0.969 ± 0.002 (0.9998)	0.9850	0.8648	1.0773	1.034 ± 0.009 (0.9985)

is introduced in the time series, except through the selection of q. We show the results for these three catalogues in Tables 23.2, 23.3, 23.4. The values of the R/S and DEA analysis show the correlation in brackets as well. Note that no summation is made for calculating the DEA exponent, but the ε^q_n is calculated for each event and the new series is analyzed.

The first three H estimators are consistent (except for $q = 0$, which will be treated later) for the Iberian Peninsula and southern California. However, the wavelet-based method provides higher values in all the cases. This is likely due to level shifts in the signal, which introduces a strong bias in this estimator. These level shifts are a consequence of the changes in the accuracy of the networks, resulting in a higher number of earthquakes detected, and a greater number of lower energy events recorded. For the energy, the H value found with the R/S analysis is in agreement with that obtained by (Ogata and Abe 1991). The R/S estimation for the Benioff strain and the approximate source dimension, however, present $H > 1/2$. The value of H increases with decreasing q. The Hurst exponent is related to the smoothness of the profile, so that a smoother graphic would provide a higher H. With $q < 1$ in Eq. (4) the effect is that of smoothing the energy, acting as a low pass filter.

We also tested the behaviour for the rate of earthquake generation, by dividing the time in 50 intervals, and calculating the number of earthquakes in each one. For

Table 23.3. Results for Greece.

q	R/S	DSOD	WDSOD	Wavelet	DEA
1	0.557 ± 0.003 (0.9992)	0.5012	0.5007	1.9364	0.124 ± 0.010 (0.9000)
1/2	0.808 ± 0.005 (0.9992)	0.5145	0.5163	1.2320	0.77 ± 0.04 (0.9438)
1/3	0.936 ± 0.003 (0.9998)	0.5294	0.5277	1.0633	0.977 ± 0.017 (0.9946)
0	1.0123 ± 0.0015 (0.9999)	-0.143	-0.0007	1.000	0.794 ± 0.011 (0.9964)
rate	0.9336 ± 0.0023 (0.9998)	0.6233	0.4555	1.0486	0.798 ± 0.010 (0.9969)

Table 23.4. Results for Southern California.

q	R/S	DSOD	WDSOD	Wavelet	DEA
1	0.580 ± 0.023 (0.9610)	0.5001	0.5004	2.6494	0.033 ± 0.003 (0.8954)
1/2	0.547 ± 0.022 (0.9586)	0.5112	0.5116	1.4873	0.200 ± 0.016 (0.8935)
1/3	0.572 ± 0.018 (0.9749)	0.5483	0.5484	1.2107	0.82 ± 0.04 (0.9641)
0	0.920 ± 0.011 (0.9966)	-0.0051	-0.0001	1.000	0.839 ± 0.004 (0.9996)
rate	0.9632 ± 0.0023 (0.9998)	0.4115	-0.2788	1.1245	0.839 ± 0.004 (0.9996)

this new time series, we calculated the Hurst and DEA exponents (last line in the tables). In general, δ increases from around 0 to 0.8 with decreasing q.

For comparison, in Tables 23.5, 23.6, 23.7 we show the values obtained for δ in each case, and the values we would obtain by supposing first a Gaussian behaviour (Eq. 1) and second a Lévy walk (Eq.2), using the Hurst exponent obtained with the R/S method. There is not a clear trend that allows us to choose either the Gaussian or the Lévy approach. In addition, as stated above (see Table 23.1), with high H the errors are higher in calculating the δ value.

The number of events is a special case. Because this is a point process, the prediction for the next value is obvious: the next number of earthquakes is 1. For this reason the wavelet-based method estimates a value H = 1. In this case, the other two MATLAB estimators do not follow this behaviour, but they found H ≈ 0, which "predicts" the range and the variance are directly proportional (see Eq. (3)).

Table 23.5. Comparison between δ from the H obtained (Eq. 1 if the time series would follow a Gaussian scheme and Eq. 2 if it is a Lévy walk) and those from DEA for the Iberian Peninsula catalogue.

q	δ_{th} Gaussian	δ_{th} Lévy	δ from DEA
1	0.51	0.51	0.04
1/2	0.63	0.58	0.30
1/3	0.79	0.70	0.87
0	0.96	0.92	0.88
rate	0.97	0.94	1.03

Table 23.6. Comparison between δ from the H obtained (Eq. 1 for Gaussian and Eq. 2 for Lévy walk) and those from DEA for the Greek catalogue.

q	δ_{th} Gaussian	δ_{th} Lévy	δ from DEA
1	0.56	0.53	0.12
1/2	0.81	0.72	0.77
1/3	0.94	0.89	0.98
0	1.01	1.03	0.79
rate	0.93	0.88	0.80

Table 23.7. Comparison between δ from the H obtained (Eq. 1 for Gaussian and Eq. 2 for Lévy walk) and those from DEA for the Southern Californian catalogue.

q	δ_{th} Gaussian	δ_{th} Lévy	δ from DEA
1	0.58	0.54	0.03
1/2	0.55	0.52	0.20
1/3	0.57	0.54	0.82
0	0.92	0.86	0.84
rate	0.96	0.93	0.84

5 Conclusions

The Hurst exponent has been calculated for the Iberian, southern California and Greek catalogues by means of four different techniques. The Hurst exponent increases from 0.5 to 1 with decreasing q. The DEA has been tested for FBM with different H. The diffusion exponent δ increases from around 0 to 0.8 with decreasing q.

No definitive FBM or Lévy-walk trend is found, most likely because of either non-stationarities inherent to the processes involved, or because of the noise

introduced by network changes. The model usually adopted to describe the time distribution of earthquakes is the Generalized Poisson (GP) model which assumes that earthquakes are grouped into temporal clusters of events and that these clusters are uncorrelated. But the results for δ with q=1/3, q=0, and the rate of the events clearly indicate that the statistics on the time distribution of these quantities is different from those of the GP model, for which δ=0.5. Perhaps, as suggested by Scafetta and West (2004) earthquake clusters are closest to 1/f long-range correlated, and present a pink noise distribution. This technique can be tuned for further analysis, which might lead to a better understanding of the underlying physics of seismic behaviour.

Acknowledgments

This work was partially supported by the MCYT project CGL2005-05500-C02-02/BTE, the MCYT project CGL2005-04541-C03-03/BTE, and the Research Group 'Geofísica Aplicada' RNM194 (Universidad de Almería, España) belonging to the Junta de Andalucía. The work of KFT was funded by an NSERC Discovery Grant. The work of AJ was founded by the 'Fundación Ramón Areces' Grant.

References

Abry, P., Flandrin, P., Taqqu, M., and Veitch, D. (2003) Self-similarity and long-range dependence through the wavelet lens, *Theory and applications of long-range dependence*, pp. 527–556, Birkhäuser.

Bowman, D. D. and Sammis, C. G. (2004) Intermittent Criticality and the Gutenberg-Richter Distribution, Pure Appl. geophys., 161, 1945–1956.

Burton, P. W., , Xu, Y., Qin, C., Tselentis, G., and Sokos, E. (2004a) A catalogue of seismicity in Greece and the adjacent areas for the twentieth century, Tectonophysics, 390, 117–127.

Burton, P. W., Qin, C., Tselentis, G., and Sokos, E. (2004b) Extreme earthquake and earthquake perceptibility study in Greece and its surrounding area, Nat. Hazards, 32, 277–312.

Carpinteri, A. and Pugno, N. (2005) Are scaling laws on strength of solids related to mechanics or to geometry?, Nature Materials, 4, 421–423.

Gimeno, R. (2000) Análisis Caótico de Series Temporales Financieras de Alta Frecuencia. El Contrato de Futuro sobre el Bono Nocional a 10 años, Ph.D. thesis, Universidad Pontificia de Comillas, Madrid.

Goltz, C. (1997) *Fractal and chaotic properties of earthquakes*, Springer-Verlag, Berlin.

Gutenberg, B. and Richter, C. (1956) Earthquake magnitude, intensity, energy, and acceleration, Bull. Seism. Soc. Am., 46, 105–145.

Hurst, H. E. (1951) Long-term storage capacity of reservoirs, Trans. Am. Soc. Civil Eng., 116, 770–808.

Hurst, H. E. (1956) Methods of Using Long-Term Storage in Reservoirs, Proc. Inst. Civil Eng., 1, 519–543.

Istas, J. and Lang, G. (1994) Quadratic variations and estimation of the local Hölder index of a Gaussian process, Ann. Inst. Poincaré, 33, 407–436.

Jiménez, A., Tiampo, K. F., Levin, S., and Posadas, A. M. (2006) Testing the persistence in earthquake catalogues: The Iberian Peninsula, Europhys. Lett., 73 (2), 171–177.

Mai, P. M. and Beroza, G. C. (2000) Source Scaling Properties from Finite-Fault-Rupture Models, Bull. Seism. Soc. Am., 90, 604–615.
Main, I. G. (1999) Applicability of time-to-failure analysis to accelerated strain before earthquakes and volcanic eruptions, Geophys. J. Int., 139, F1–F6.
Makropoulos, K. C. and Burton, P. W. (1985) A catalogue of seismicity in Greece and adjacent areas, Geophys. J. R. Astron. Soc., 65, 741–762.
Makropoulos, K. C., Drakopoulos, J. K., and Latousakis, J. B. (1989) A revised and extended earthquake catalogue for Greece since 1900, Geophys. J. Int. Res. Note, 98, 391–394.
Mandelbrot, B. and Wallis, J. R. (1968) Noah, Joseph and the operational hydrology, Water Resour. Res., 4, 909–918,.
Mathworks: MATLAB, http://www.mathworks.com/.
Ogata, Y. and Abe, K. (1991) Some statistical features of the long-term variation of the global and regional seismicity, Int. Stat. Review, 59, 139–161.
Papazachos, B. C. and Comninakis, P. E. (1971) Geophysical and tectonic features of the Aegean arc, J. Geophys. Res., 76, 8517–8533.
Papazachos, B. C. and Papazachou, C. (1997) *The Earthquakes of Greece*, Editions Ziti, Thessaloniki.
Papazachos, B. C., Karakaisis, G. F., and Hatzidimitriou, P. M. (1994) Further information on the transform fault of the Ionian sea, *XXIV Gen. Ass., Europ. Seism. Com., Athens, 1924 September*, 1, 377–384.
Piron, P. (2005) Transform techniques and non-stationarity with an emphasis on network applications, Master's thesis, University of Liège, Belgium.
Sammis, C. G. and Smith, S. W. (1999) Seismic Cycles and the Evolution of Stress Correlation in Cellular Automaton Models of Finite Fault Networks, Pure Appl. Geophys., 155, 307–334.
Scafetta, N., Hamilton, P., and Grigolini, P. (2001) The thermodynamics of social processes: the teen birth phenomenon, Fractals, 9, 193–208.
Scafetta, N. and West, B. J. (2004) Multiscaling Comparative Analysis of Time Series and a Discussion on 'Earthquake Conversations' in California, Phys. Rev. Lett., 92, 138 501–138 504.
Scafetta, N. and West, B. J. (2005) Multiscaling comparative analysis of time series and geophysical phenomena, Complexity, 10 (4), 51-56.

24 Snow Avalanches as a Non-critical, Punctuated Equilibrium System

D.M. McClung

Department of Geography, University of British Columbia, Vancouver, B.C. V6T 1Z2 Canada, mcclung@geog.ubc.ca, Tel: 604-822-9157, Fax: 604-822-6150

Abstract. The mathematical requirements for a self-organized critical system include scale invariance both with respect to the characteristic sizes of the events and the power spectrum in the frequency domain based on time arrival. Sandpile avalanches and other types of avalanches have been analyzed from the perspective of common characteristics of critical systems. However, snow avalanches have not been completely analyzed, particularly in the frequency domain. Snow avalanches constitute a natural hazard and they are of much more practical importance than other types of avalanches so far analyzed. In this chapter, I consider the mathematical criteria for scale invariance in both the size and frequency domain for snow avalanches based entirely on analysis of field measurements.

In combination, the mathematical results suggest that neither the size distribution, the time arrival nor waiting time between avalanche events conform to that of a critical system as defined for self-organized criticality or thermodynamics. If snow avalanches are to conform to a critical system in geophysics then a revision of the mathematical requirements or definition is called for. However, time series of events show that snow avalanche arrivals consist of clusters, intermittancies and bursts with rapid changes over short time intervals interrupted by periods of stasis. The data and analysis combined with field observations suggest that a system of snow avalanches paths exhibits the characteristics of a non-critical, punctuated equilibrium system.

1 Introduction

Bak, Tang and Wiesenfeld (1987) provided the original description of a self-organized critical system. They defined self-organized criticality (SOC) for complex systems as implying that no single characteristic event size or time scale exists. The essential meaning is that there is not just one time scale or length scale that controls the temporal or spatial evolution of such a system. According to Jensen (1998): “For a critical system, the same perturbation applied at different positions or at the same position at different times can lead to a response of any size.” The original application in earth geophysical systems was applied to sandpile avalanches by a hypothetical computer model. Subsequently, it was shown that SOC does not apply to sandpile avalanches when they were observed instead of relying on computer modelling. Since the term “avalanche” is used quite often in description of SOC, it is

a fair question to ask about applicability to snow avalanches since they occur frequently and they have important effects on human life causing death and destruction. If SOC applies (or if it does not apply), then important implications for avalanche forecasting are implied. In common with the analysis of sandpile avalanches, field observations and measurements are used in this chapter instead of theoretical conjecture or computer models to test if SOC applies to snow avalanches.

This chapter is divided into two general parts to check the mathematical requirements for self-organized criticality to apply. The first part deals with examination of scale invariance in relation to sizes of snow avalanches and the second part deals with time arrival and waiting time to estimate the power spectrum in frequency domain. The application to snow avalanches is of some importance since snow avalanches along with earthquakes constitute two of a very few geophysical systems with a high enough frequency of events to perform analysis for both the size and frequency (or time) domain. Landslides, debris flows and rock avalanches, for example, do not have a high enough frequency of events for analysis in the time/frequency domain.

Most claims about the applicability of SOC (Bak 1996) involve checks on either scale invariance of the power spectrum in the frequency domain (power law dependence) or scale invariance with respect to event sizes (power law behaviour of some quantity which represents size). However, Bak, Tang et al. (1987) and Jensen (1998) emphasize that scale invariance in both the frequency and size domain are needed to demonstrate SOC. In this chapter, I suggest through use of field data from snow avalanches that neither of these mathematical requirements are met. However, some of the descriptive aspects of SOC appear to have merit based on field observations and measurements including large catastrophes and bursts of activity over short time scales which are characteristic of punctuated equilibrium systems.

2 Loose Snow Avalanches: Size Information

Loose snow avalanches form [see (McClung and Schaerer 2006) for details on loose snow avalanches] in cohesion-less snow a the surface and are analogous to sandpile avalanches. When they occur (see Fig. 24.1), they are generally of nearly the same size. I suggest that this is due to inertial effects. Even though they may initiate with different sizes, the final size is determined by inertial effects and they sweep out snow of similar size (volume). There are no data to show this effect in general. However, field observations suggest it is a fairly general trend and Fig. 24.1 supports the idea. Without a spectrum of sizes then, I suggest that loose snow avalanches share a characteristic with sandpile avalanches which violates SOC. Namely that final sizes are controlled by inertial effects and implicitly SOC should not apply since the final masses are nearly all the same size during times when the physical condition of the surface layers remains roughly the same.

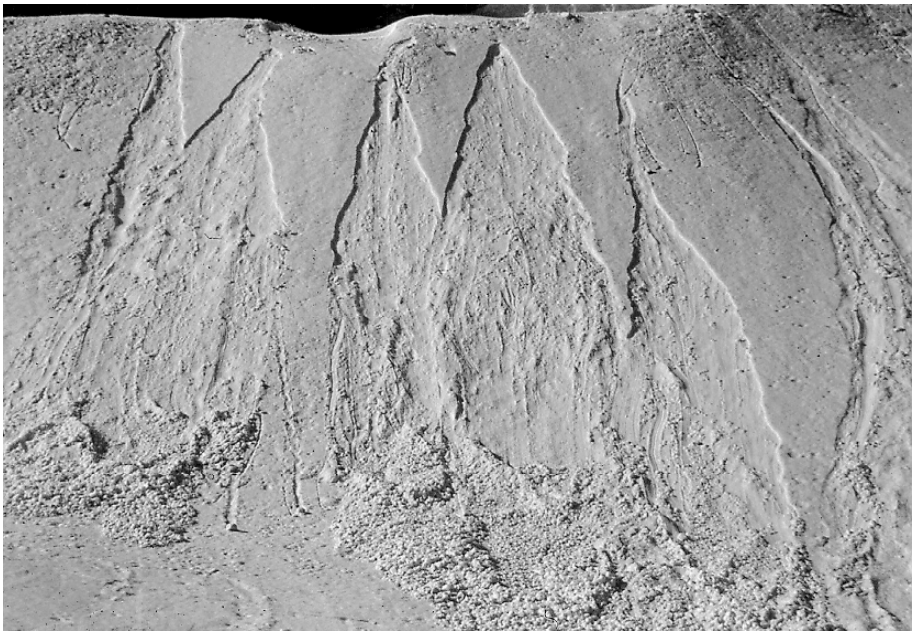


Fig. 24.1. Loose snow avalanches. Photo: SLF archives. This example shows they are all nearly the same size.

3 Scale Invariance with Respect to Multiplicative Changes in Slab Thickness

Slab avalanches (Figure 24.2) initiate by shear fracture propagation (McClung, 1979; 2005) within a thin weak layer at a depth D (the slab thickness) below the snow surface (see Fig. 24.3). In general, they mostly release due to snow loading during storms or from loading by blowing snow (see (McClung and Schaerer 2006) for an extensive review of slab avalanche formation). The depth to the weak layer, D, is similar for most slab avalanches that release in storms in an area, with the result that there is a characteristic length scale associated with release. Thus, when they occur in an area, they are mostly of the same size. The characteristic length scale for release violates the principles of SOC, so I suggest that it is highly unlikely that SOC applies to slab avalanche release. No analysis is needed to make this claim but field observations and measurements confirm the existence of the thin weak layer and its role in the introduction of a length scale (D) which is involved in slab release.

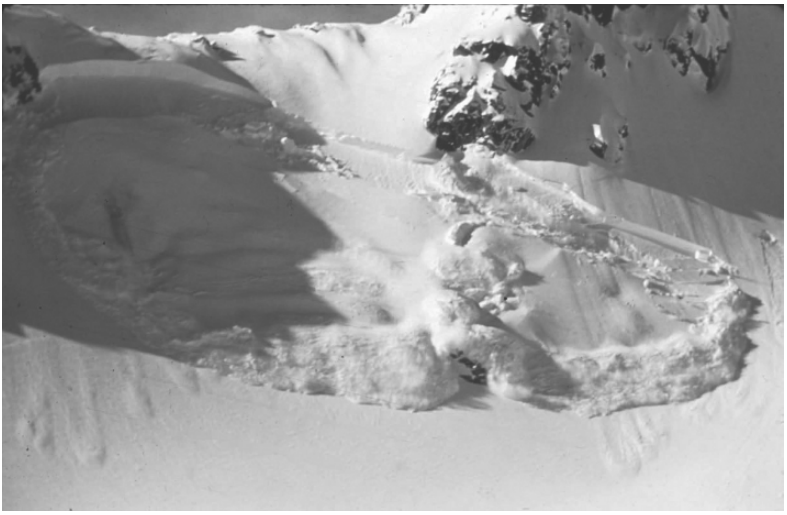


Fig. 24.2. Example of slab avalanche release , Whistler Bowl, British Columbia. Photo by T. Salway. The avalanche releases by shear fracture at depth D under the slab.

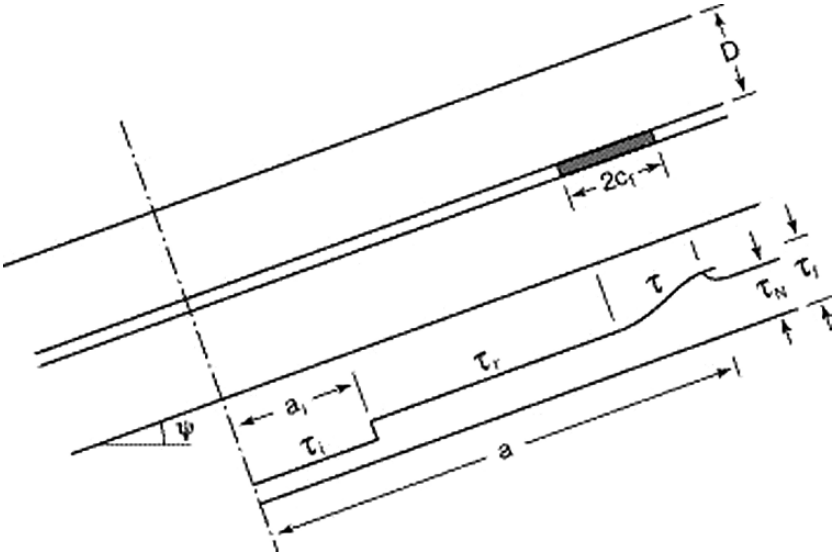


Fig. 24.3. Schematic for release of a snow slab on a slope of angle ψ after Bažant et al. [2003]. Strain-softening occurs in shear with the fracture process zone $(2c_f)$ of finite length in a weak layer at depth D. Failure initiates at an imperfection of length a_i with shear stress τ_i in the weak layer. The weak layer has failure stress τ_f, residual stress, τ_r and τ_N is applied nominal shear stress.

It is still of interest to analyse the distribution of sizes, particularly, the fundamental size parameter D. Figure 24.4 shows a probability plot of D for 191 slab avalanches. The data suggest that the fundamental length scale, D, follows a log-normal probability density function. The data in Figure 24.4 are from mix of avalanche triggers including: snow loading (called naturals), skier triggering and explosive loading. I have analysed (McClung 2003 a) other data sets from skier triggering and natural loading and the results are similar. In each case, D is log-normally distributed. Since the data range only from about 0.1 m to about 4 meters, it is not possible that the data follow a power law or scale invariant form (Pareto distribution) over any significant range of D (e.g several orders of magnitude). I have also made similar plots for slab length and slab width and they also display log-normal form. Thus, I conclude that the application of SOC fails with respect to size for slab avalanches for two reasons: 1. there is an implicit length scale: D, the slab

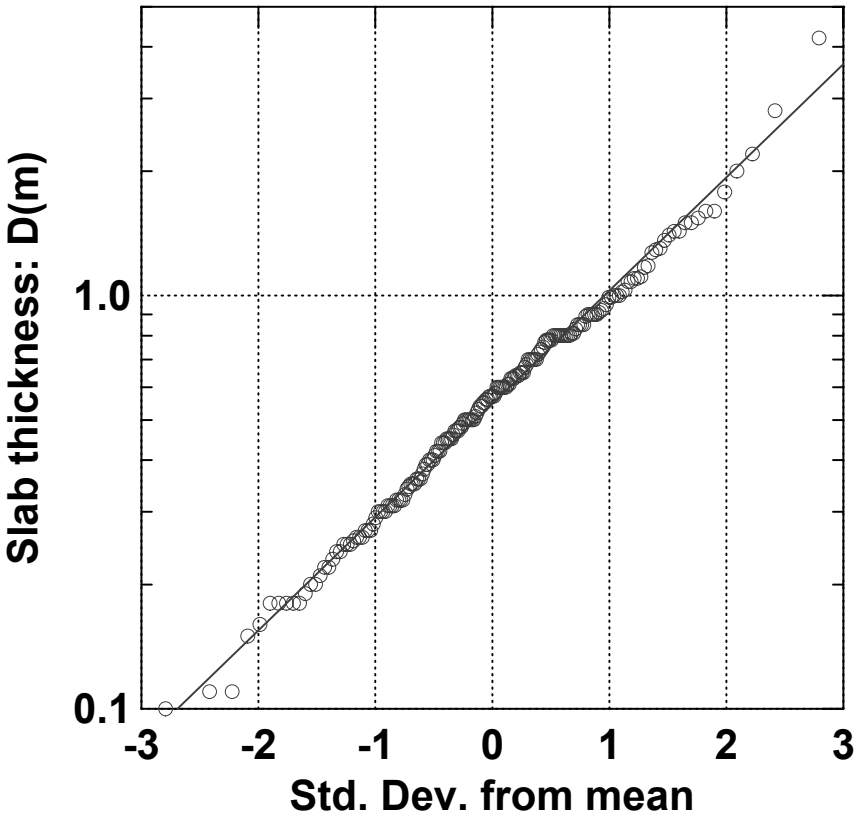


Fig. 24.4. Probability plot of slab avalanche thickness D (m). The plot suggests D follows a log-normal probability density function. The abscissa represents the z score in standard deviations from the mean.

thickness associated with release and this violates the fundamental assumptions of SOC; 2. The actual sizes do not display scale invariant (power law) behaviour over any significant range of the size scale D or other size characteristics such as slab length and width. I have discussed this in detail (McClung 2003 a).

4 Analysis in the Time and Frequency Domain for Slab Avalanches

Snow slab avalanches, along with earthquakes, constitute two geophysical systems with a high enough frequency of events to enable analysis for scale invariance in the frequency or time domain to check the assumptions required for a self-organized critical system. Bak et al. (1987) suggested that a self-organized critical system should display scale invariance of the power spectrum in the frequency domain with proportionality to : $1/f$ or $1/f^{\beta}$ where β is a constant that is usually between one and two. If such dependence was proven then no characteristic time scale would be available. For avalanches, this could affect the fundamental properties of avalanche forecasting. The idea that geophysical systems might have some simple common properties is also appealing. Since the ideas about self-organized criticality started with analysis of sandpile avalanches, it is useful to check the assumptions with full scale snow avalanche events as a natural geophysical system. In addition to checking the assumptions of self-organized criticality, such an analysis should provide fundamental information about time scales, whether they are present or not, in relation to avalanche forecasting.

McClung (2003(b)) provided an empirical analysis of avalanche time arrival from Bear Pass and Kootenay Pass in British Columbia from thousands of avalanches based on 20 years of records. The analysis revealed that a time scale of about 1 day appears from the times series analysis of arrivals which does not conform to the scale invariance in time postulated for a self-organized critical system. This results of the analysis are briefly reviewed in the next section (5). The reader is referred to McClung (2003(b)) for all the details of the analysis.

In this chapter, another analysis is performed from data bases of avalanche arrivals for Bear Pass and Kootenay Pass. The analysis here is done from the perspective of an analysis of waiting time between snow avalanches to check for $1/f$ dependence in the frequency power spectrum. The average waiting time between events is calculated for 110 different avalanche paths from 15,990 avalanches over 23 winter seasons in which approximate time of occurrence is known. The waiting time analysis is given in sections 6-9 below.

5 Review of Empirical Arrival Time Analysis for Bear Pass and Kootenay Pass

McClung (2003(b)) calculated time arrival for avalanche events for determination of the frequency power spectrum for Bear Pass and Kootenay Pass. Direct calculation from the time series of events collected in 12 hour intervals showed that the auto-correlation function had exponential form. Each element of the series consists of sum of the masses of all avalanches recorded within 12 hour intervals. The sizes of the avalanches according to the Canadian avalanche size classification (McClung and Schaerer 2006) were recorded by technicians and these were then converted to mass with all masses summed over 12 hour intervals to yield time series for each pass. See McClung (2003(b)) for more details.

The auto-correlation function for such a time series is an empirical relation derived by correlating the time series with itself displaced by one or more lags. By definition, the auto-correlation function is 1 at zero lag since it represents the series correlated with itself. In this manner, the auto-correlation function is equivalent to an empirical histogram of the time series correlated with itself for lags 0,1,2,.....where in this case one lag equals 12 hours. The empirical form derived from the time series of events (McClung 2003(b)) after the constant (d.c.) component was removed is:

$$c(t,\tau_0) = \exp(-t/\tau_0) \tag{1}$$

where $\tau_0 \approx 1-2$ days (the e-folding time) for both Bear Pass and Kootenay Pass was shown by (*McClung*, 2003(b)). Figures 24.5 shows the time record of total avalanche mass (t) summed for all avalanche paths at Bear Pass for 20 years of records from 1981 -2001 with data stratified into 12 hour time periods where 12 hours constitutes a lag of one. Summer periods from April 30 - November 1 were removed. It shows that the time record of summed avalanche arrivals (all events are naturals without human triggering) consists of a series of clusters, intermittancies and bursts. Figure 24.6 shows the empirical auto-correlation plot obtained directly from the time series from (McClung 2003(b)). Similar plots are available for Kootenay Pass (*McClung,* 2003(b)). The frequency power spectrum associated with equation (1) is:

$$S(f,\tau_0) = 4\,\mathrm{Re}\int_0^\infty c(t,\tau_0)e^{2\pi ift}dt = 4\tau_0/[1+(2\pi f\tau_0)^2] \tag{2}$$

This analysis suggests a finite time scale ($\tau_0 \approx 1-2$ days) associated with the avalanche arrivals with transition to white noise for low frequencies. Typical storm durations generating avalanches can last from a fraction of a day to as long as about 10 days in the avalanche areas considered here. These empirical results suggest that scale invariant behaviour in the frequency power spectrum for time arrival is not expected over any significant range of frequency or time. McClung (2003(b)) provides further details to generalize this result for the system modelled as a shot noise Poisson process.

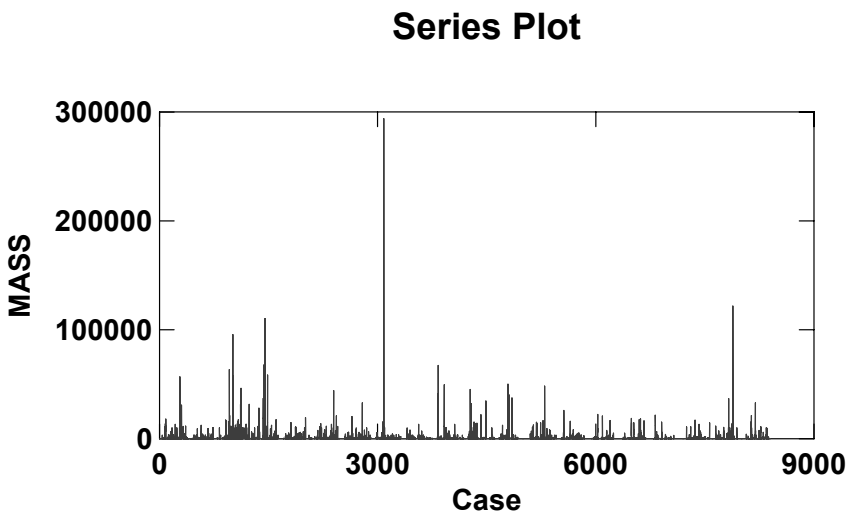


Fig. 24.5. Time series of summed avalanche mass (tonnes) for all the avalanche paths at Bear Pass, British Columbia for data from 1981 -2000. Each case consists of a 12 hour period (one lag). Summer periods were removed to provide an approximately stationary time series. The data consist of natural events unaffected by human avalanche control.

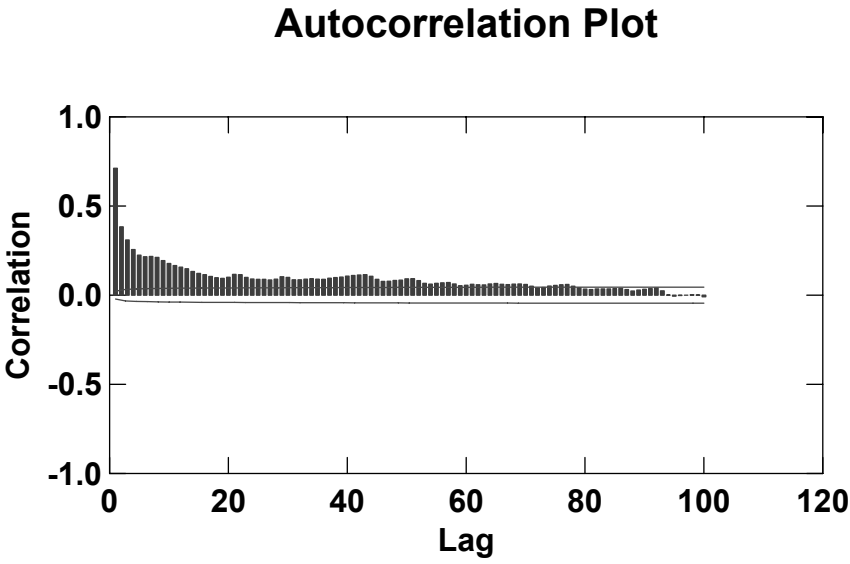


Fig. 24.6. First order auto-correlation function for time arrival of events for 20 years of records from Bear Pass, from (McClung 2003(a)) for the time series of Figure 24.5. The horizontal lines represent the limits of two standard errors (or 95% confidence limits): correlations outside inside these lines are not significant. One lag equals 12 hours.

Conclusions about frequencies higher than the Nyquist frequency (once a day) cannot be made for the analysis since the effective sampling rate is twice per day (12 hour intervals or lags). McClung (2003(b)) found by non-linear fits to the auto-correlation function that the exponential decay of the auto-correlation function as in equation (1) provided a better fit than a half-Gaussian function for both Bear Pass and Kootenay Pass. I suggest the plots in Figures 24.4 and 24.5 illustrate the after-effects of random perturbations applied by storms to generate avalanche arrivals for the systems of avalanche paths at Bear Pass and Kootenay Pass with decline in correlation to white noise (constant power spectrum) over a limited interval in a time of several and perhaps up to ten τ_0.

6 Waiting Time Analysis as a Poisson Process

The results above suggest that there is a time scale associated with summed avalanche arrivals when the collection of avalanche paths is viewed as a system. This is in agreement with field observations and data that almost all avalanches arrive during or within a day or so after storms. The time series analysis suggests that an avalanche area viewed as a system has short time memory and the after effects of random perturbations (storms in this case), correlation declines as $\exp[-t/\tau_0]$ with the power spectrum as in equation (4). Clearly this does not represent scale-free invariance in the frequency domain as envisioned for SOC by Bak et al. (1987).

Another possibility for analysis in the time/frequency domain is to analyse waiting time between avalanches. Ito (1995) suggested scale invariance in the frequency power spectrum by an analysis of earthquake waiting times in California. Thus, it is of interest whether such behaviour might apply to snow avalanches. In this section, I performed the analysis for large data bases of avalanche occurrences based on 23 years of records from Bear Pass and Kootenay Pass. Below, I attempt to analyse correlations in the time/frequency domain by decomposing the signal into a linear random superposition of independent components. The components consist of individual, independent avalanche paths.

I assume (McClung 1999) that avalanches for a given avalanche path arrive as random, rare, discrete independent events such that the probability of two events arriving simultaneously is negligible. These conditions fit avalanche arrivals and they conform to the assumptions of a Poisson process. If the random variable T is the time to the first arrival, then the probability that T exceeds some value of time t is equal to the probability that no events occur in t. The probability of exceedance is: $1-P_i(t)$ where $P_i(t)$ is the cumulative distribution function for avalanche path, i. For Poisson arrivals, the probability of no events in time t is:

$$e^{\frac{-t}{\tau_i}}\left(\frac{t}{\tau_i}\right)^0\frac{1}{0!}=1-P_i(t) \qquad t\geq 0 \tag{3}$$

Where τ_i is the average waiting time between events for avalanche path i. From equation (3), the probability density function for path i is (Benjamin and Cornell 1970):

$$\psi_i(t) = \frac{dP_i(t)}{dt} = \frac{1}{\tau_i}\exp\left(\frac{-t}{\tau_i}\right) \tag{4}$$

Since for a Poisson process there is no memory of past events, the analysis applies for any interval of time whether or not it begins at $t = 0$. Thus, if time is counted from the last event (nth), then $\exp(-t/\tau_i)$ is the exceedance probability or the probability that the time to the $(n+1)th$ event is greater than t. It is important to distinguish between the constant τ_i (the average waiting time between events for an individual path i) and the empirically found τ_0 in the analysis above which refers to time arrivals for the entire avalanche areas at Bear Pass and Kootenay Pass. The latter reflects time arrivals and it is closely related to a characteristic relaxation time for the system of avalanche paths as a whole.

7 Description of Data for Waiting Time Analysis

The data used in the analysis consist of avalanche events recorded at two avalanche areas in British Columbia (B.C.), Canada. The times of avalanche events were recorded between November 1, 1981 and April 30, 2005 for sets of avalanche paths at Bear Pass in northwest B.C. and Kootenay Pass in southeast B.C. The data base consists of 15,990 avalanches recorded for these areas for 23 winter seasons from November 1 each year until April 30 each year. Data from all avalanche paths for these two areas were included except at Bear Pass two paths were excluded which had very high frequency of events with multiple gullies spawning small avalanches so they were not distinct paths. Also, events recorded as loose snow avalanches were excluded from the analysis so that only slab avalanches were considered.

Since the snow and avalanche season is generally from November 1 - April 30, events outside these dates were excluded. Thus, the data sets can be envisioned as having a stationary continuous flow of events over the time scale used (1981 - 2005) with the summer time periods removed as described by McClung (2003(b)) for the arrival analysis above. There are occasional events before November 1 and after April 30 but generally they are not recorded and omission of these few events should not affect the analysis given here.

Average waiting times between events were calculated for each avalanche path by determining the total number of avalanches recorded over the 23 season time period and the total time between the first and last events recorded. Avalanche paths with

only one event over the entire 23 years were excluded since it was not possible to determine a value for the time between events. The average time between events for an avalanche path was determined by calculating the time range over which events were recorded and then the number of summer periods through the time range were subtracted. This total time was then divided by : $N_0 - 1$ where N_0 is the total number of events recorded for a path. The data set included 5855 avalanches from 37 paths at Kootenay Pass and 10,135 avalanches from 73 paths at Bear Pass. Table 24.1 contains descriptive statistics for the two avalanche areas with mean and standard deviation of waiting times for the avalanche areas given.

Table 24.1. Descriptive statistics for Bear Pass (B. Pass) and Kootenay Pass (K. Pass) with mean and standard deviations ($\overline{\tau_i}, \sigma_{\tau_i}$) of waiting time given for avalanches (Aval.).

Area	No. Aval.	No. Paths	$\overline{\tau_i}$	σ_{τ_i}
B. Pass	10135	73	43 days	59 days
K. Pass	5855	37	94 days	134 days

Figures 24.7 and 24.8 contain probability plots of mean waiting times for the avalanche paths. In both cases, the ordinate is logarithmic so that the plots suggest the waiting times follow log-normal probability density functions. Table 24.2 contains relevant descriptive statistics.

Table 24.2. Mean and variance of $\ln(\tau_i)$ for Bear Pass (B. Pass) and Kootenay Pass (K. Pass) from 15,990 avalanches recorded over 23 winter seasons.

Area	No. Paths	$\overline{\ln(\tau_i)}$	$\sigma^2_{\ln(\tau_i)}$
B Pass	73	3.1	1.1
K Pass	37	3.5	2.3

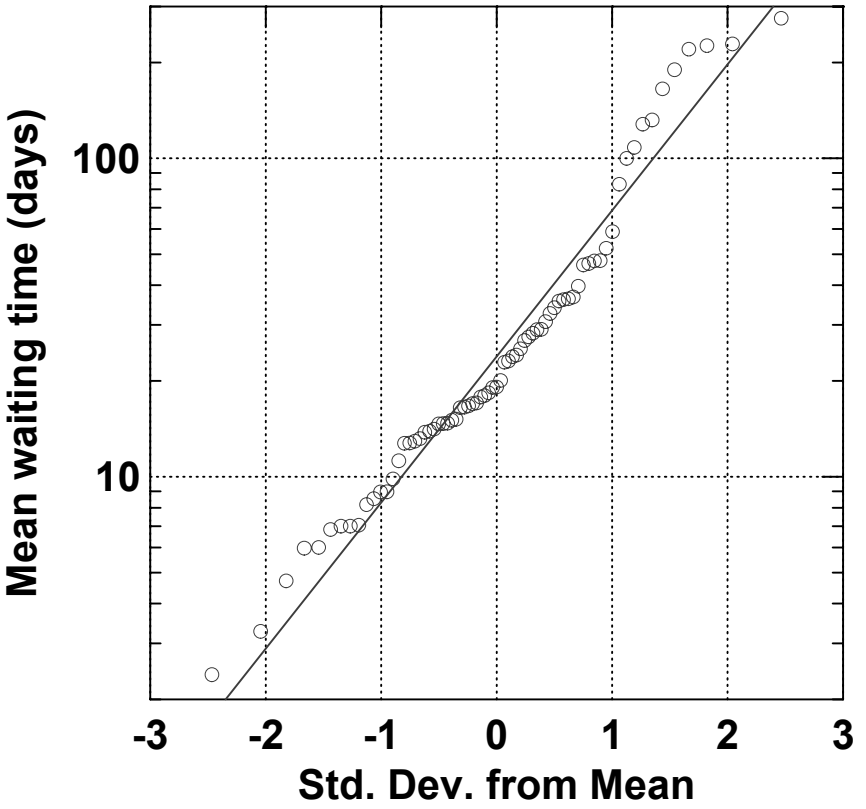


Fig. 24.7. Probability plot of mean waiting time for 73 avalanche paths from Bear Pass. The plot suggests that mean waiting time has a log-normal probability density function for the avalanche area. The abscissa represents the z score in standard deviations from the mean.

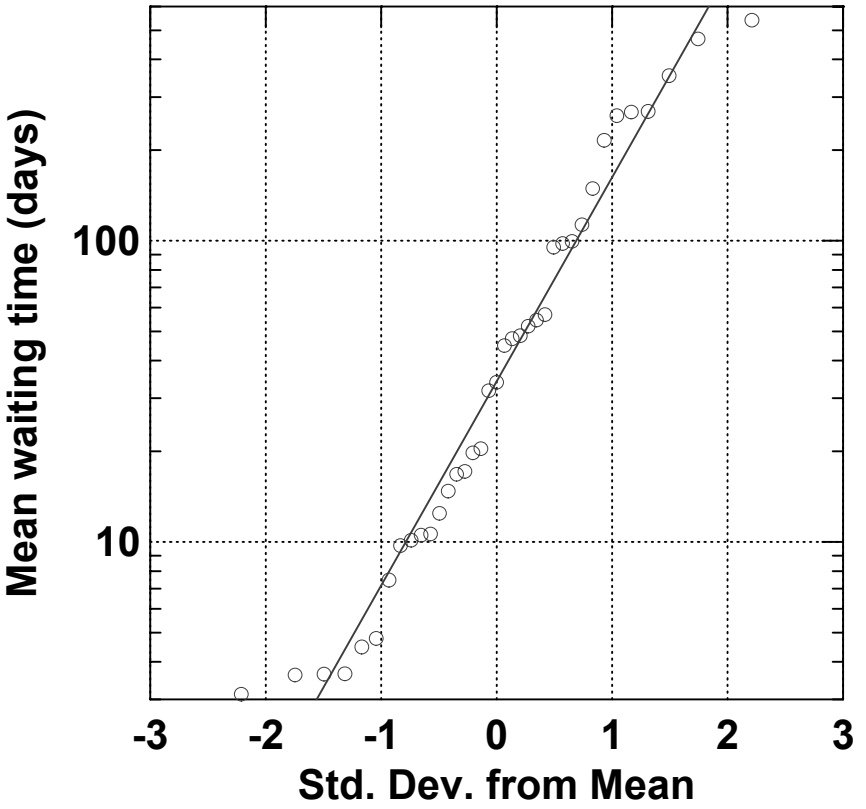


Fig. 24.8. Probability plot of mean waiting time for 37 avalanche paths from Kootenay Pass similar to Fig. 24.7. The plot suggests that mean waiting time has a log-normal probability density function. The abscissa represents the z score in standard deviations from the mean.

8 Effects of Avalanche Control

The avalanche paths at Kootenay Pass (K. Pass) are heavily controlled with gunfire much more so than at Bear Pass (B. Pass). In order to assess the effects of avalanche control, the analysis above was repeated for natural occurrences only. The results are given in Tables 24.3 and 24.4 which may be compared with Tables 24.1 and 24.2.

Table 24.3. Descriptive statistics for natural events (Aval.) only for Bear Pass (B. Pass) and Kootenay Pass (K. Pass) with mean and standard deviations $(\overline{\tau}_i, \sigma_{\tau_i})$ given.

Area	No. Aval.	No. Paths	$\overline{\tau_i}$	σ_{τ_i}
B. Pass	8550	73	54 days	82 days
K. Pass	1225	35	96 days	135 days

Table 24.4. Mean and variance of $\ln(\tau_i)$ for 9775 naturally occurring avalanches recorded over 23 seasons for Bear Pass (B. Pass) and Kootenay Pass (K. Pass).

Area	No. Paths	$\overline{\ln(\tau_i)}$	$\sigma^2_{\ln(\tau_i)}$
B Pass	73	3.3	1.1
K Pass	35	4.5	1.0

Comparison of the tables shows that far fewer avalanches are naturals for Kootenay Pass and the variance $\sigma^2_{\ln(\tau_i)}$ is much lower for Kootenay Pass when only natural events are considered from Kootenay Pass. The variance is comparable to Bear Pass if explosive controlled events are eliminated. Probability plots for natural events only are very similar to Figures 24.3 and 24.4 and they imply that the collection of mean waiting times τ_i follows a log-normal distribution for either avalanche area.

9 Rare Events Approximation for Waiting Time Analysis

The analysis in this paper is done from the perspective to the rare events approximation (McCormick 1981) applied to the avalanche areas for avalanche waiting times. The rare events approximation consists of formulations for the intersection of events and the union of events applied to sets of avalanche paths at Bear Pass and Kootenay Pass. The basic assumption is that occurrences are independent and highly infrequent. It is assumed that the probability of two or more events is small compared to the probability for any single event. These assumptions will apply to events on a given avalanche path. For the entire avalanche areas consisting of 73 paths (Bear Pass) and 37 paths (Kootenay Pass), I have found (see Appendix A) that the union of events for the rare events approximation holds over only a very short restricted time scale (order of a few hours) whereas the intersection

of events should be valid without restrictions since all the avalanche paths are independent. Consider an avalanche path A_i and let $P(A_i)$ be the probability of an event on this path. For paths $1, N$, the union of events is then:

$$P(A_1 \cup A_2 \cup ...A_N) \approx \sum_{i=1}^{N} P(A_i) \tag{5}$$

and the intersection is given by:

$$P(A_1 \cap A_2 \cap .....A_N) = P(A_1)P(A_2).....P(A_N) \tag{6}$$

Equations (5) and (6) together constitute the rare events approximation. The results for the union of events are derived in Appendix A from which it is inferred that self-similar temporal scaling (temporal scale invariance) is only a useful concept for individual avalanche paths with very long waiting times between events. For the avalanche areas as a whole, the time scale of applicability is on the order of a few hours.

In this section, the formulation for the intersection of events (equation 6) is developed based on independence of the individual paths. The object is to assess whether the frequency power spectrum has $1/f$ dependence over any significant frequency range for the collection of avalanche paths in the two areas.

Assuming completely random events, the probability density function for an avalanche path takes the form of equation (4) and the auto-correlation function is easily calculated as:

$$c(t, \tau_i) = \exp(-t/\tau_i) \tag{7}$$

For a single avalanche path, the frequency power spectrum, $S(f)$, is of the same form as equation (2) with the replacement : $\tau_i \rightarrow \tau_0$. From Figs. 24.7 and 24.8, the analysis suggests that the probability density function for τ_i in both areas is log-normal. Thus, a distribution of mean waiting times is assumed such that:

$$\int_0^{\infty} \rho(\tau) d\tau = 1 \tag{8}$$

In equation (8) the subscript on τ is dropped and a continuous distribution of waiting times $\rho(\tau)$ is assumed.

Since the individual avalanche paths are separate and disjoint, I assume no correlations between them and the power spectrum for an avalanche area is given by the weighted sum from the individual avalanche paths (*Van der* Ziel 1950; Montroll

and Shlesinger 1983; Schroeder 1990; Jensen 1998). From equation (2), the power spectrum for an avalanche area then becomes:

$$S(f) = \int_0^\infty \frac{4\tau\rho(\tau)\, d\tau}{[1+(2\pi f\tau)^2]} \tag{9}$$

From Figs. 24.7 and 24.8, I suggest that is reasonable to consider $\rho(\tau)$ as log-normal. One may also derive this by considering $\tau = 1/P$ as the inverse of a probability per unit time following *Montroll and Shlesinger* (1983). Assuming all the τ_i are independent as in equation (6), then

$$\log(\tau) = \log(1/P) = \sum_{i=1}^{N} \log(1/p_i) = \sum_{i=1}^{N} \log(\tau_i) \tag{10}$$

In equation (10), P represents the probability per unit time that a 'grand success' occurs meaning that the system of avalanche paths is represented in series such that a 'grand success' requires that all are active. Since each p_i is a random variable then so is τ_i and equation (10) follows from the independence assumption of equation (6). From the central limit theorem if N becomes large, and the first and second moments exist, then it is implied that $\log(\tau)$ has a normal distribution or τ has a log-normal one, as suggested by the data. Now let $\overline{\tau}$ be the mean value of τ . From Table 24.1, the values of $\overline{\tau}$ are 43 days for Bear Pass and 94 days for Kootenay Pass or 54 days and 96 days from Table 24.3 if only natural events are considered.

The log-normal distribution has the form:

$$\rho\left(\frac{\tau}{\overline{\tau}}\right) d\tau = \frac{\exp\{-[\log(\tau/\overline{\tau})]^2 / 2\sigma^2_{\ln(\tau)}\}}{(2\pi\sigma^2_{\ln(\tau)})^{1/2}} \frac{d\tau/\overline{\tau}}{\tau/\overline{\tau}} \tag{11}$$

The power spectrum is then:

$$S(f) = 4\overline{\tau}\int_0^\infty \frac{(\tau/\overline{\tau})\exp(-[\log(\tau/\overline{\tau})]^2 / 2\sigma^2_{\ln(\tau)}) d(\tau/\overline{\tau})}{(2\pi\sigma^2_{\ln(\tau)})^{1/2}(\tau/\overline{\tau})[1+(2\pi f\tau)^2]} \tag{12}$$

Some important limits from the integration of (12) were derived by Montroll and Shlesinger (1983).

If $\sigma^2_{\ln(\tau)} \to \infty$ then

$$S(f) \sim \frac{4\overline{\tau}}{4\pi f \overline{\tau} \sigma_{\ln(\tau)}} \left(\frac{\pi}{2} \right)^{1/2} \tag{13}$$

and a $1/f$ spectrum results.

If $f \to 0$ then

$$S(f) \sim 4\overline{\tau} \exp\left(\frac{1}{2} \sigma^2{}_{\ln(\tau)} \right) \tag{14}$$

and white noise (constant power spectrum) results at low frequency similar to the low frequency limit of the power spectrum for an individual path as in equation (2).

If $f \to \infty$ then

$$S(f) \sim \frac{4\overline{\tau}}{(2\pi f \overline{\tau})^2} \exp\left(\frac{1}{2} \sigma^2{}_{\ln(\tau)} \right) \tag{15}$$

and Brownian noise ($1/f^2$) spectrum results similar to the high frequency limit for a single path (equation 2).

These results (equations 13-15) are related to those of Van der Ziel (1950), Schroeder (1990) and Jensen (1998). They all assumed (on an ad-hoc basis) a wide (high variance) uniform distribution of τ to produce a $1/f$ spectrum over a wide range of frequencies with $1/f^2$ behaviour at high frequencies and white noise at low frequencies.

Since the spectrum of equation (12) goes from white noise at low frequency to $1/f^2$ at high frequency it is suggested that there may be a range of intermediate frequencies in which the spectrum mimics $1/f$ noise (also called flicker noise or pink noise). The other possibility for $1/f$ noise arises if $\sigma^2{}_{\ln(\tau)}$ is very large. However, for the present analysis, the values in Tables 24.2 and 24.4 show that such is not the case.

Montroll and Shlesinger (1983) performed the asymptotic integration of the power spectrum given by equation (12). If $\left|\ln(2\pi f \overline{\tau})\right| < 2\sigma^2{}_{\ln(\tau)}$ then $S(f)$ has a simplified log-normal form:

$$S(f) = \frac{C_0}{\sigma_{\ln(\tau)} \pi f} \left(\frac{\pi}{2} \right)^{1/2} \exp\left[-\frac{1}{2} \left(\frac{\ln\{2\pi f \overline{\tau}\}}{\sigma_{\ln(\tau)}} \right)^2 \right] \tag{16}$$

where

$$C_0 = \left(\left[1 - \frac{1}{2}\left(\frac{\pi}{2\sigma_{\ln(\tau)}} \right)^2 + \frac{5}{8}\left(\frac{\pi}{2\sigma_{\ln(\tau)}} \right)^4 \right] + \ldots \right).$$

From Tables 24.1 and 24.2, the condition that equation (16) applies is: $f < 0.033 = 1/30(day)^{-1}$ (Bear Pass) and $f < 0.169 = 1/5.9(day)^{-1}$ (Kootenay Pass). These calculations show the importance of the mean waiting time, $\overline{\tau}$, and the variance, $\sigma^2_{\ln(\tau)}$, in order for the spectrum to have the simple log-normal form of equation (16). Kootenay Pass has higher mean and variance to greater possibility for the log-normal form. For natural events only, the above values are replaced by $f < 0.027 = 1/37(day)^{-1}$ (Bear Pass) and $f < 0.012 = 1/83(day)^{-1}$ (Kootenay Pass) and the effect of the reduced variance considering only natural events at Kootenay Pass is evident.

From the form of equation (16), it is clear that if the argument of the exponential is $<<1$ then the spectrum approximates $1/f$ dependence. Following Montroll and Shlesinger (1983) taking the log of equation (16) gives a form which can be analysed for approximating $1/f$ dependence:

$$\ln[S(f)] = -\ln(f) - \frac{1}{2}[(\ln f)^2 / \sigma^2_{\ln(\tau)}] - \ln(C_1) \tag{17}$$

where C_1 is a constant. If the spectrum is to be $1/f$ then the middle term must be much smaller than the first term so that only the first and last terms are present. Let $f = \exp(n_0)$ then equation (17) becomes:

$$\ln[S(f)] = -n_0 - \frac{1}{2}(n_0{}^2 / \sigma^2_{\ln(\tau)}) - \ln(C_1) \tag{18}$$

If the second term is to be less than a small fraction, θ, of the first term (Montroll and Shlesinger 1983) then

$$|n_0| \le 2\theta\sigma^2_{\ln(\tau)} \tag{19}$$

and $S(f)$ approximates a $1/f$ distribution for $4\theta\sigma^2_{\ln(\tau)} + 1$ e-folds to within a relative error θ. If $\theta = 0.1$ then, within relative error of 10% and with values from Table 24.2, this amount to 1.4 e-folds for Bear Pass and 1.9 e-folds for Kootenay Pass. If only natural events are considered from Table 24.4, the spectra from both areas follow $1/f$ with a 10% error for about 1.4 e-folds.

Generally, to claim scale invariance ($1/f$ dependence in this case) for a critical geophysical system (Jensen 1998; Korvin 1992) it must be demonstrated at least two and preferably several orders of magnitude. From the analysis here, I conclude that this condition is not fulfilled. The principal reason is that the variance, $\sigma^2_{\ln(\tau)}$, is not large enough for either system of avalanche paths at Bear Pass and Kootenay Pass. For the avalanche areas here, the waiting times for individual paths vary only by about two orders of magnitude. From equation (17), it is clear that if $\sigma^2_{\ln(\tau)}$ is large enough then $1/f$ dependence results and it is the only physical parameter that matters for the present analysis once the frequency requirement for satisfying equation (16) is attained.

10 Discussion and Conclusions

The postulates of self-organized criticality include lack of both characteristic length and time scales for a series of events. I have shown previously (McClung 2003(a)) that for snow slab avalanches, there is always a characteristic length scale associated with the events, namely, the depth to the weak layer determines a characteristic size and the avalanches are nearly all the same size during any short time frame. In contrast, for the definition of self-organized criticality provided by Bak *et al. (*1987) a spectrum of sizes is expected. In addition, even if the characteristic sizes of slab avalanches are analysed (McClung 2003(a)) they do not follow a power law distribution over any significant range of sizes as suggested for SOC. The limited range of slab thickness combined with the fact that the probability plots of slab depth, length and width all suggest they follow log-normal probability density functions suggests that scale invariance with respect to size is not followed.

Snow avalanches and earthquakes constitute two of a very few entities in solid earth surface geophysics with a high enough frequency of events for analysis of time dependence for testing the definition of a critical system. For the mountain slope hazards (debris flows, landslides, rock avalanches, snow avalanches, rock falls), snow avalanches and rock falls (due to high frequency of events) are the only ones with a reasonable chance of testing the postulated scale invariance of the power spectrum in the frequency domain required for an SOC system. A major object of the present chapter has been to provide a fairly comprehensive test of whether or not scale invariance exists in the frequency domain for snow avalanche events. Previously, (McClung 2003(b)), I analysed time arrival data for Bear Pass and Kootenay Pass, and I found consistent results that a characteristic time scale is present. In this chapter, a companion analysis of waiting time between avalanches is provided. The results of the present analysis are consistent with the previous results. Namely, it is unlikely that frequency scale invariance is present in the waiting time analysis over any significant time or frequency range. The results with respect to waiting time here rest on the assumption that avalanches can be regarded as Poisson

events (McClung 1999) so any conclusions can be regarded as within the framework of model assumptions. Thus, if the assumptions of the model are incorrect, then so will the conclusions be incorrect or inconclusive. Table 24.5 contains a summary of the mathematical results of this chapter in comparison with the postulates of SOC based on the analysis presented.

Table 24.5. Postulates of SOC compared with results from field measurements and observations of snow avalanches.

Postulate of SOC	Snow avalanche results
No characteristic size	Loose snow avalanches are nearly all the same size for the same snow pack conditions.
No characteristic size or scale invariant size distribution	Slab avalanches have a characteristic size (or length scale) D. The cumulative distribution for D is not scale invariant over any significant range.
Scale invariant frequency power spectrum (no characteristic time scale)	Slab avalanche arrivals viewed as a system in an avalanche area have a characteristic relaxation time τ_0.
Scale invariant frequency power spectrum (e.g $\sim 1/f$)	Slab avalanche waiting time may have power spectrum approximately proportional to $1/f$ for about 1 - 2 e-folds with 10% error.

When the results of the present analysis are combined with my previous results, I suggest that snow avalanches (both loose and slab avalanches) do not conform to the mathematical formalism about self-organized criticality postulated by Bak et al. (1987). For slab avalanches neither with size nor frequency invariance applies.

In spite violation of the mathematical postulates of SOC to snow avalanches and other related phenomena such as sandpile avalanches, it is an undeniable achievement of Per Bak to attempt to highlight the possible important commonalities of geophysical events such as the mountain slope hazards and earthquakes. There is definitely something critical about the release of a snow slab avalanche, a rock avalanche, a flake type (e.g. quick clay) landslide or an earthquake caused by fracture propagation along rock joints. They all exhibit slow build-up and evolution until catastrophic, rapid failure occurs (sometimes under very small influences). For example, dry snow slabs sometimes are observed when air temperatures change without additional mechanical load being applied. Further, the mountain slope hazards and earthquakes, all may have strain-softening behaviour (see Figure 24.3) during failure and prior to fracture which suggests they can be regarded as due to failure of quasi-brittle materials with implied fracture mechanical size effects and important length scales prior to self-propagating fractures as outlined by Bažant and Planas (1998) and Bažant et al. (2003) for the snow slab.

The presence of a time scale in snow avalanche arrivals or waiting times is consistent with human knowledge in relation to avalanche forecasting. Most natural avalanches happen during or immediately after snow storms with failure on weak layers in the storm snow. Persistent weak layers such as surface hoar, faceted snow or depth hoar [see McClung and Schaerer (2006) for more information] can result in large avalanches sometimes months after the weak layers have formed. However, upon application of perturbations of snow loading by storms or blowing snow, nearly all natural avalanches happen during or immediately (for example within a day or two) after load application.

The definition of a time scale (e.g. return period) or waiting time for individual avalanche paths is the basis for land-use planning in avalanche terrain. It depends importantly on terrain characteristics and the snow climate (McClung 2003(c)). Of course, for either a return period or a wait time, there is no information about 'when' events will occur. Only the statistical properties of many avalanches including their size and frequency give rough ideas about how often, on average, events are expected and expected sizes. *Bak* [1996] emphasizes this statistical or probabilistic character as an important characteristic of a critical system.

The presence of a time scale in arrivals or wait times for natural, open, non-equilibrium systems is probably more the rule (Korvin 1992) than the exception. However, Bak et al. (1987) postulated for SOC that lack of a time scale is necessary. A more relaxed criterion for the frequency power spectrum, $S(f)$, such as a decay from Brownian noise at high frequency to white noise at low frequency might be a more suitable common property for prospective geophysical critical systems in nature. A scale invariant form such as $1/f$ or $1/f^{\beta}$ seems to require a very high variance in waiting time to evolve, particularly if scale invariance is to be present over a significant frequency range such as more than one order of magnitude. Since it is not possible (Van der Ziel 1950; Montroll and Schlesinger 1983) for a form like $1/f$ to be present over all frequencies (e.g. the spectrum diverges at zero frequency) then natural variations in frequency must vary over at least 3, and perhaps 4, orders of magnitude in order to achieve scale invariance over two orders of magnitude. For snow avalanches and the mountain slope hazards, such wide variations are not easily observed or verified.

From observations, there are two distinct cases that characterize time instability (McClung 2002(b)) for an avalanche area which I suggest may be classed as an open, non-equilibrium system. Most of the time, conditional instability prevails, whereby additional energy must be supplied to produce fractures and trigger avalanches. This situation produces occasional avalanches at isolated locations which are difficult to locate. The difficulty in locating potential fracture points is the primary source of fatal accidents in back-country travel which constitute more than 90% of avalanche deaths in North America and Europe (McClung and Schaerer 2006). The second type is: absolute instability, whereby the system reaches critical instability in many avalanche paths over considerable spatial extent (sometimes over an entire mountain range) and there are widespread natural releases. The latter

produces a 'burst' of activity, called an avalanche cycle. Within the time frame of a winter, conditional instability is the rule (prevailing most of the winter) and absolute instability is the exception, often being observed for only a few hours in an avalanche area during a winter (McClung 2002(b)).

The time scale of several hours for self-similar clustering of wait times (several hours) for events at Bear Pass and Kootenay Pass (derived in Appendix A) matches the observed time scale for absolute instability suggested by McClung (2002(b)). Rare winters with high losses and many large avalanches, such as 1999 in the European Alps of Switzerland, Austria and France are observed when a persistent weak layer forms early in the winter (in this case over nearly the entire Alps) with overloading by heavy snow falls later to generate large avalanches from the deeply buried weak layer over a storm time scale of one to several days. In this sense, the two cases of instability for snow avalanches provide at least a qualitative match to Per Bak's (Bak 1996, p. 59 - 62] description of a non-equilibrium, open system and the expected character of events. The absolute instability situation corresponds to evolution of the system making large transformations (avalanche cycles) over short time scales (hours for avalanche areas; days for a mountain range) as Bak also suggests for self-organized criticality. In contrast, the conditional instability represents a relative 'stasis' with isolated events in time and space. The clustering of events separated by periods of stasis is analogous to punctuated equilibrium described by Bak (1996) as a characteristic of complex non-equilibrium systems some of which exhibit $1/f$ noise and fractal behaviour in relation to event size to comprise SOC. There is no argument about the fact that a system of snow avalanche paths displays the signals of punctuated equilibrium. However, the analysis in this chapter combined with field observations suggests that the $1/f$ noise and fractal behaviour are absent.

Large transformations over short time scales are occasional features of scientific thought and Per Bak may have been an architect of such a transformation. For snow avalanches, the mathematical details suggest snow avalanches are non-critical with respect to the original ideas about SOC. However, human experience supported by field measurements suggests the physical ideas do have merit. In the midst of an avalanche cycle, the notion of an average waiting time or return period has very limited relevance. Instead, we can expect Bak's time-scale-free dependence to prevail. In time, avalanche events (see Figure 24.5) consist of clusters, intermittancies and bursts (Birkeland and Landry 2002; McClung 2003(b)). Avalanche forecasting (McClung 2002 (a)) is concerned with the specifics of predicting rapid changes in space and time relative to a given triggering energy level so forecasters implicitly follow the principles of punctuated equilibrium which is one of descriptive aspects of SOC . However, such forecasting was developed on the basis of practical experience long before SOC was proposed. The unstable equilibrium which characterizes snow slab avalanche release (Figure 24.2) suggests a different kind of criticality which is highly significant to humans and which is shared by earthquake, rock avalanche and rapid landslide release. The prediction problem reduces to much the same thing for either type of criticality: statistical and probabilistic rather than

deterministic. However, the details of criticality for the snow slab are more likely rooted in fracture mechanics (McClung 1979) instead of scale invariance.

Acknowledgements

This research was funded by Canadian Mountain Holidays and the Natural Sciences and Engineering Research Council of Canada. I am very grateful for this support. I am also thankful for data from Bear and Kootenay Passes which were provided by the British Columbia Ministry of Transportation and Highways. The suggestion to analyse wait times was provided by Professor Garry Clarke. The mathematical brilliance of Elliott Montroll made the wait time analysis possible.

References

Bak, P. (1996), How nature works, the science of self-organized criticality, *Springer-Verlag New York Inc.*, New York, U.S.A., 212 pp.

Bak, P., Tang, C. and K. Wiesenfeld (1987), Self-organized criticality: An explanation of $1/f$ noise, *Physical Review Letters, 59*, 381 - 384.

Bažant, Z.P. and J. Planas (1998), Fracture and size effect in concrete and other quasibrittle materials, *CRC Press*, Boca Raton, USA, 616 pp.

Bažant, Z.P., G. Zi and D.McClung (2003), Size effect and fracture mechanics of the triggering of dry snow slab avalanches, *J. Geophys. Res., Vol. 108*, *No. B2,* 2119, doi:10.1029/ 2002JB001884.

Birkeland, K.W. and C.C. Landry (2002), Power laws and snow avalanches, *Geophys. Res. Lett, 29(11)*, 49-1–49-3.

Benjamin, J.R. and C. A. Cornell (1970), Probability, statistics and decision for civil engineers*, McGraw- Hill Inc*, New York, U.S.A., 684 pp.

Ito, K. (1995), Punctuated equilibrium model of biological evolution is also a self-organized critical model of earthquakes, *Phys. Rev. E* (52): 3232.

Jensen, H.J. (1998) Self-organized criticality, emergent complex behaviour in physical and biological systems, Cambridge Lecture Notes in Physics 10, *Cambridge University Press*, Cambridge, U.K., 153 pp.

Korvin, G. (1992), Fractal models in the earth sciences, *Elsevier*, Amsterdam, 396 pp.

McClung, D.M. (1979), Shear fracture precipitated by strain-softening as a mechanism of dry slab avalanche release, *J. Geophys. Res.*, *84(B7)*, 3519-3526.

McClung, D.M. (1999), The encounter probability for mountain slope hazards, *Canadian Geotechnical Journal*, *Vol. 36, No. 6,* 1195 -1196.

McClung, D.M. (2002(a)), The elements of applied avalanche forecasting part I: the human issues, *Natural Hazards* , *Vol. 26, No. 2,* 111-129.

McClung, D.M. (2002(b)), The elements of applied avalanche forecasting part II: the physical issues and rules of applied avalanche forecasting, *Natural Hazards*, *Vol. 26, No. 2,* 131-146.

McClung, D.M. (2003(a)) Size scaling for dry snow slab release, *J. Geophys. Res.*, *Vol. 108, No. B10,* 2465, doi:1029/2002JB002298.

McClung, D.M. (2003 (b)) Time arrival of slab avalanche masses, *J. Geophys. Res*. *Vol. 108, No. B10,* 2466, doi:10.1029/2002JB002299.

McClung, D.M. (2003(c)) Magnitude and frequency of avalanches in relation to terrain and forest cover, *Arctic, Antarctic and Alpine Res. Vol. 35, No.1*, 82-90.

McClung, D.M. (2005), Dry slab avalanche shear fracture properties from field measurements, *J. Geophys.Res*., *110*, F04005, doi:10.1029/2005JF000291.

McClung, D. and P. Schaerer (2006), The Avalanche Handbook, 3rd Edition, *The Mountaineers Books*, Seattle, Wash., 342 pp.

McCormick, N.J. (1981), Reliability and risk analysis, methods and nuclear power applications, *Academic Press, Inc*., Boston, U.S.A., 446 pp.

Montroll, E.W. and M.F. Shlesinger (1983), Maximum entropy formalism, fractals, scaling phenolmena, and $1/f$ noise: a tale of tails, *Journal of Statistical Physics*, *Vol. 32, No.2,* 209-230.

Schroeder, M. (1990), Fractals, chaos, power laws, minutes from an infinite paradise, *W.H. Freeman and Company*, New York, 429 pp.

Shlesinger, M.F. and B.D. Hughes (1981), Analogs of renormalization group transformations in random processes, *Physica 109A,* 597-608**.**

Van der Ziel, A. (1950), On the noise spectra of semi-conductor noise and flicker effect, *Physica XVI, no. 4,* 359-372.

Appendix A : Implication of Rare Events Approximation for a Set of Disjoint Avalanche Paths for the Union of Events

Consider a set of avalanche paths, $A_{1,}A_2....A_N$ which are disjoint with rare events assumed not to overlap in time. Under the rare events approximation (McCormick 1981;p.16): the probability of the union of events is given by:

$$P(A_1 \cup A_2 \cup A_3 \cup ....A_N) \approx \sum_{i=1}^{N} P(A_i) \tag{A1}$$

As a guideline, the rare events approximation holds when $P(A_i) \le 0.1$.

For a random process with a sequence of events, the waiting time density between events is $\psi(t)$ and $\psi(t)dt$ is the probability that an event occurs in the interval $(t, t+dt)$ with the most recent event at t = 0. Assuming the rare events approximation holds, for N avalanche paths partitioned in the sample space of an avalanche area, the probability in terms of the normalized density is:

$$\psi(t)dt = dt\sum_{i=1}^{N} a_i\psi_i(t) = \frac{dt}{N}\sum_{i=1}^{N}\psi_i(t) \tag{A2}$$

where $\sum_{i=1}^{N} a_i = 1; a_i = \frac{1}{N}.$

In Laplace transform space, for an individual path, i,

$$\psi_i^*(\varepsilon) = \int_0^{\infty} e^{-\varepsilon t}\psi_i(t)dt \tag{A3}$$

For completely random Poisson events, the waiting time between events is taken to be exponentially distributed from equation (4):

$$\psi_i(t) = \frac{1}{\tau_i}\mathrm{e}^{-\frac{t}{\tau_i}}; t \ge 0 \tag{A4}$$

where τ_i is the mean waiting time between events for avalanche path i . From (A3),

$$\psi_i^*(\varepsilon) = \frac{1}{1+\tau_i\varepsilon} \tag{A5}$$

The probability density for the time of occurrence of nth event on path i is:

$$\psi_{i,n}(t) = \int_0^\infty \psi_{i,n-1}(t-\tau)\psi_i(\tau)d\tau,$$

$$\psi^*_{i,n}(\varepsilon) = \int_0^\infty \exp(-\varepsilon t)[\psi_{i,n}(t)]dt \tag{A6}$$

For a process started at t = 0,

$$\psi^*_{i,n}(\varepsilon) = [\psi^*_i(\varepsilon)]^n \tag{A7}$$

For an individual avalanche path, from (A4) and (A6), the time to the nth event is given by the gamma distribution (Benjamin and Cornell 1970):

$$\psi_{i,n}(t) = \frac{1}{\tau_i^n (n-1)!} t^{n-1} \exp(-t/\tau_i) \tag{A8}$$

From (A5) and (A7),

$$\psi^*_{i,n}(\varepsilon) = \frac{1}{\left(1+\tau_i\varepsilon\right)^n} \tag{A9}$$

If $\tau_i\varepsilon >> 1$ (or $t/\tau_i << 1$), then the scale invariant form:

$$\psi^*_n(\varepsilon) \sim \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\left(\tau_i\varepsilon\right)^n} = \frac{1}{\varepsilon^n \overline{\tau_{inv}}^n} \tag{A10}$$

results where

$$\overline{\tau_{inv}}^n = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\tau_i^n}\right)^{-1}. \tag{A11}$$

Direct integration of (A2) gives the cumulative distribution function:

$$P_n(t) = \frac{t^n}{n!}\frac{1}{\overline{\tau_{inv}}^n}\left[1-\left(\frac{n}{n+1}\right)\frac{t}{\overline{\tau_{inv}}} + O\left(\frac{t^2}{\overline{\tau_{inv}}^2}\right) - \ldots\right] \tag{A12}$$

and if $nt << (n+1)\tau_i$, $i = 1, N$ then the scale invariant form:

$$P_n(t) = \frac{t^n}{n!\overline{\tau_{inv}}^n} \tag{A13}$$

results for the avalanche area under the rare events approximation.

For Bear Pass (73 paths) and Kootenay Pass (37 paths), values for $\overline{\tau_{inv}}$ are 15 and 14 days respectively. These results suggest that a finite time scale from exponential decay will be associated with the process unless the minimum of the average wait times is large. Given the minimum of the mean waiting times, τ_i, for individual paths at Kootenay Pass (3 days) and Bear Pass (2 days), the power law (scale invariant) behaviour implied by (A13) applies only for times less than a few hours respectively for either pass. If the rare events approximation does not hold $(t > 0.1\tau_i)$ then a finite time scale will be associated with the process but the analysis would be intractable.

For a Poisson process, there is no memory of past events so that time $t = 0$ can be taken at any arbitrary time of the most recent event. The time estimated above for scale invariant time dependence (or self-similar clustering of event wait times) for the avalanche areas is less than storm durations during which many avalanche events are observed over less than half a day. If time $t = 0$ is taken near the beginning of a storm, then there will be time overlap between events and the rare events approximation will not hold. For Bear Pass, the probability of one event in a time of less than 4 hours is about 1/76 and the probability of two events is about 1/ 2825 from (A13). These results suggest that self-similar clustering or scale invariant behaviour is of short duration for these avalanche areas. If the minimum wait time was on the order of months or years, instead of days, different conclusions would arise.

In general, self-similar clustering of event waiting times for an individual avalanche path, i, will occur if $t \ll \tau_i$ or $\tau_i \varepsilon \gg 1$. In this case, the decay of $\psi_{i,n}(t)$ will be slow enough that no finite time scale will occur in the analysis and a scale invariant form $\psi_{i,n}(t) \sim t^{n-1}$ [see equation (A8)] develops. For $t << \tau_i;\ i = 1, N$ (or $t < 0.1\,\tau_i$ approximately) the rare events approximation for the union of events is expected to apply and each term in the finite series (A1) has the scale invariant form $P_{i,n} \sim t^n$ so self-similar clustering of event waiting times arises as in (A13) for a collection of paths.

If the mean waiting time $\tau_i \to \infty$, then no characteristic time appears and self-similar clustering of event times for an avalanche path occurs. This was shown by Shlesinger and Hughes (1981) and Montroll and Shlesinger (1983) for temporal scaling of Poisson events. From (A9), if τ_i is finite, then Taylor series expansion around $\varepsilon \to 0$ (for long times) gives:

$$\psi^*_{i,n}(\varepsilon) = 1 - n\varepsilon\tau_i + \frac{1}{2!}(n)(n+1)\varepsilon^2{\tau_i}^2 - \ldots\ldots \qquad \text{(A14)}$$

Equation (A14) shows that a finite time scale will be contained in the process at long times unless the mean waiting time is infinite. For $\tau_i \to \infty$ then $\psi^*_{i,n}(\varepsilon), \psi_{i,n}(t)$ both take scale invariant form.

For the events at Bear Pass (73 paths) and Kootenay Pass (37 paths), the mean of the waiting times between events for all paths taken together are 43 and 94 days (or 54 and 96 days for natural events only) respectively. McClung (2003(a)) showed from calculation of the auto-correlation function from time series of events at Bear Pass and Kootenay Pass that an e-folding time of about 1 day is associated with arrivals which suggests a short mean time scale for event occurrence.

The differences in time scales (between the mean waiting values and the e-folding arrival time) may be related to the lack of applicability of the rare events approximation for Bear Pass and Kootenay Pass time arrival analyses of McClung (2003(a)). It would be impossible to record events exactly when they occur over many paths even though they rarely, if ever, overlap in time if the observation time is short enough. The occurrence data in the analysis of McClung (2003(a)) were lumped into 12 hour time intervals. This would imply time overlap during storms in the analysis so the rare events approximation would not apply. Instead, the highest frequency paths (minimum wait times) would control the time scale for the analysis. Thus, minimum rather than average wait times would apply and the short finite relaxation time scale suggested by the time series seems reasonable since it is comparable to minimum wait times and storm durations at Bear Pass and Kootenay Pass.

From the analysis above, I conclude that self-similar clustering of event waiting times for the avalanche areas of Bear Pass and Kootenay Pass is unlikely except for short periods. The fundamental reason is that paths of short mean waiting time are present which limits the time scale of applicability as derived from the approximations needed. In turn, this suggests that a finite time scale associated with exponential decay should be associated with the process for any longer time duration. The short time 'bursts' of avalanche activity associated with large avalanche cycles which occur about once a year in avalanche areas may be related a lack of waiting time scale for short periods. I have termed such short time 'bursts' absolute instability (McClung 2002 (b)).

For individual avalanche paths, the average delay between events on paths ranges up to about 540 days at Kootenay Pass (for all events and natural events) and 275 days at Bear Pass (385 days for natural events). For avalanche paths with such long delay times, self-similar clustering of event times (or scale invariant behaviour) can be expected over fairly substantial times which exceed storm durations.

25 Evidence from Wavelet Lag Coherence for Negligible Solar Forcing of Climate at Multi-year and Decadal Periods

John Moore,[1] Aslak Grinsted[1,2] and Svetlana Jevrejeva[3]

[1] Arctic Centre, University of Lapland, 96101 Rovaniemi, Finland, jmoore@ulapland.fi
[2] Department of Geophysics, P.O. Box 3000, FIN-90014, University of Oulu, Finland ag@glaciology.net
[3] Proudman Oceanographic Laboratory, Liverpool, UK, sveta@pol.ac.uk

Abstract. We examine possible links between solar cycle irradiance variations the large atmospheric circulation systems that affect whole planet's climate. In particular we examine the putative mechanism of solar forcing mediated by changes in induced stratospheric conditions over the polar regions. We test this hypothesis by examining causal links between time series of solar irradiance based on both amplitude and length of the 11-year solar sunspot cycle and indices of Arctic Oscillation AO and ENSO activity. We use a wavelet lag coherence method based on wavelet filtering to examine the significance and magnitude of the phase coherence of the pairs of series in lag-period space. Hence we study the non-linear phase dynamics of weakly interacting oscillating systems. The method clearly shows no link between AO or SOI with solar irradiance at all scales from biannual to decadal. We conclude that the 11-year cycle sometimes seen in climate proxy records is unlikely to be driven by solar forcing.

1 Introduction

Decadal cycles are fairly ubiquitous across the planet, and are therefore persuasive of a global-scale climate mechanism (Jevrejeva, Moore and Grinsted, 2004; Moron, Vautard and Ghil 1998; Dijkstra and Ghil 2005). Several authors have been tempted to ascribe to solar sunspot cycles, however, detailed statistical analysis of many of these correlations shows them to be spurious or statistically insignificant (Laut,2003; Tsiropoula 2003). The 11-12 year solar sunspot cycle produces rather weak (0.1%) changes in solar energy output, and this is unlikely to directly to be sufficient to produce changes in weather and climate. Amplification factors have been proposed due to the higher variability of solar energy at UV wavelengths which may induce changes in stratospheric ozone and temperature, which can then propagate down to the troposphere (e.g. Baldwin and Dunkerton 2005; Labitzke 2005).

The main features of the planet's climate are the ENSO and the polar annular modes. The strength of the polar stratospheric vortex determines the index of annular mode, which are called the Arctic Oscillation, (AO) and the Antarctic Annular Mode (AAM) (Thompson and Wallace 1998). Almost all plausible sun-climate links rely

on modification of the polar stratosphere. Though Hedfors, Aldahan, Kulan, Possnert, Karlsson, and Vintersved (2006) in common with many others discuss the link between cloud formation and cosmic ray intensity – when sunspot numbers are low, more cosmic rays reach lower altitudes and the atmosphere becomes cloudier. Laut (2003) showed how problematic these studies were, and it is further shown in the later study where the authors point out that there is a large lag between ^{7}Be (a proxy for cloudiness) and the solar cycle of up to 1.5 years, and the dataset extends over only 1 solar cycle. Physically we may expect cloud condensation nuclei to be active for very short times, and so there should be phase delays very close to zero.

Kuroda and Shibata, (2005) modeled the impact of solar cycle on the AAM using a coupled chemistry-climate model in two 21-year long model runs with constantly repeating Sea Surface Temperature (SST). They found that increased ultra-violet radiation led to a more persistent signal from the AAM in the Antarctic stratosphere than during low UV model runs due to formation of an ozone anomaly (amounting to 2-3%). Furthermore they show that it is UV rather than cosmic rays that produce the difference in their model.

Barnston and Livezey (1989), and later Hameed and Lee (2005) showed that stratospheric perturbations are more likely to penetrate to the troposphere during solar cycle maximum than minima, and that the effect is also dependent on the direction of the zonal wind direction in the tropics. However these analyses rely on only data available from 1948 and hence are statistically rather insignificant. Kodera and Kuroda (2002) interpreted re-analyses data from 1979 to 1998 and proposed a mechanism for the dynamical and radiative forcing of the stratosphere by the solar cycle, while the analysis is provocative, there must be doubt to its statistical robustness as less than two whole solar cycles are included in the data set. While it is clear that stratospheric anomalies can penetrate downwards to the troposphere, it is a rather atypical phenomenon (Baldwin and Dunkerton 1999; 2001), and in general the troposphere drives the stratosphere. However, it is clear that from both observational and modeling studies that the stratosphere can provide an efficient and fast transport mechanism for linking tropical and polar climate (Baldwin and Dunkerton 2005; Jevrejeva et al. 2004), thus the stratosphere provides a bridge between the annular modes and ENSO phenomena, and so we may expect it be one factor that it is especially sensitive to the solar cycle.

Moore, Grinsted and Jevrejeva (2006) found that sunspot number is not significant factors in climate on multi-year and decadal timescales. They analysed causality relationships using wavelet coherence methods, which are developed further in this paper, and a new method of representing phase relationships is introduced. Wavelet coherence is useful as relative phase relationships between two time series across a wide spectrum of temporal scales are produced. If the variable represented by one of the time series is really the causal agent of the variability in the second time series, then a change in the first must always precede a reaction in the second. We will discuss not the sunspot numbers here but the solar radiation received at the Earth's surface, which has only been measured globally since the satellite era, but which is extended backwards in time in two distinctly different ways: based on the length of the sunspot cycle, and on the intensity of the sunspot cycle.

2 Methods and data

We use monthly time series of the AO (Thompson and Wallace 1998.), spanning 1899-2001. ENSO time series comes from monthly SOI (Ropelewski and Jones 1987) spanning 1865-2005. Moore et al. (2006) used the monthly International Sunspot numbers as the measure of the solar cycle (http://sidc.oma.be/DATA/monthssn.dat) As accurate measurements of total and surface solar irradiance variations have been made for only 2-3 decades, reconstructions based on sunspot number are needed prior to 1978. There are several different available, but here we use two that have been recently compared with climate statistics by Solanki and Krivova (2003) back to 1700; before 1978 these are reconstructions following Fligge and Solanki (2000). A secular increase in the total solar irradiance of about 2 Wm^{-2} since the Maunder minimum is assumed, TSI_A follows the solar cycle amplitude evolution, and TSI_L follows solar cycle length evolution. We removed the mean monthly values (the annual cycle) from all series.

The method we use determines the non-linear interactions between the two time series that may be chaotic. We extract the phase expression of the time series derived from the Continuous Wavelet Transform (CWT) of a time series (e.g. Grinsted, Moore and Jevrejeva 2004; Torrence and Compo 1998). The idea behind the CWT is to apply the wavelet as a band pass filter to the time series. As we desire a broad band pass filter, we use the Paul as this is not very localized in frequency space, and allows signals that are relatively aperiodic to be included in the analysis:

$$\psi_0(\eta) = \frac{2^m i^m !}{\sqrt{\pi(2m)!}}(1 - i\eta)^{-(m+1)} \tag{1}$$

where ω_0 is dimensionless frequency and η is dimensionless time, and m is the order, taken as 4 here. The centre frequency of the Paul wavelet, λ, is an important parameter in the analysis and is given by

$$\lambda = (2m+1)/(4\pi s)). \tag{2}$$

The wavelet is stretched in time by varying its scale (s), so that $\eta = s \cdot t$, and normalizing it to have unit energy. The CWT of a time series X, $\{x_n, n=1,\ldots,N\}$ with uniform time steps δt, is defined as the convolution of x_n with the scaled and normalized wavelet.

$$W_X(s,t)\big|_{t=n} = \sqrt{\tfrac{\delta t}{s}} \sum_{n'=1}^{N} x_n \psi_0\left[(n'-n)\tfrac{\delta t}{s}\right]. \tag{3}$$

The complex argument of $W_X(s,t)$ can be interpreted as the phases of $X\{\varphi_1\ldots, \varphi_N\}$ at the scale s. We utilize the angle strength of the phase angle difference between two series (X and Y), also known as the mean phase coherence, $\rho(X,Y)$ (Mokhov and Smirnov 2006). We are interested in causative relations, so it is appropriate to measure ρ between the phases φ_x, φ_y of the two time series.

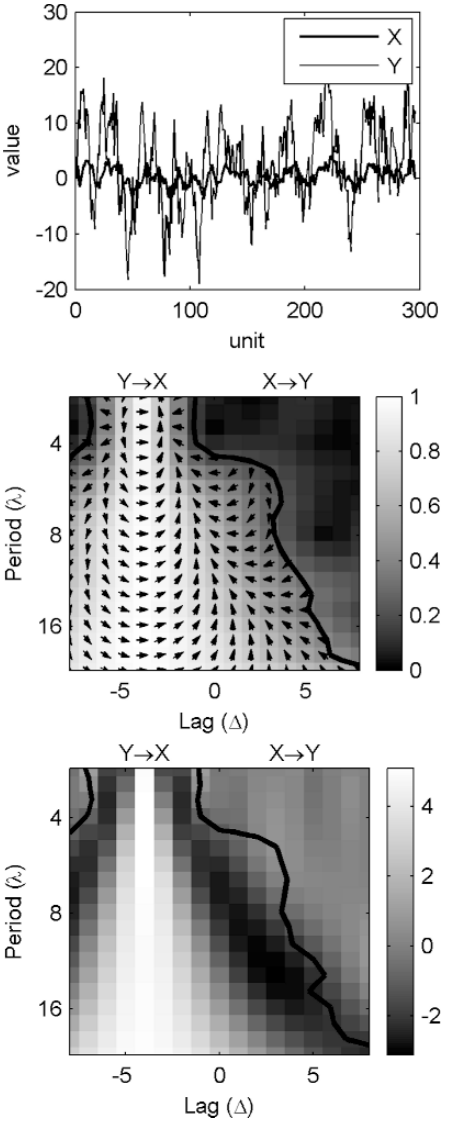


Fig. 25.1. Top: Plot of X and Y data, a simple linear regression gives a best fit of data: $Y = 1.9X + 0.61$, which fails to capture any important linkage between the two noisy series. Middle: the wavelet lag coherence plot showing values of mean phase coherence (ρ) and its 95% confidence interval by the solid black contour. Note the arrows points to the right at a lag of -4 indicating that is when X and Y are in phase at all λ. Bottom: the sensitivity, m, in the equation $Y = mX$ for the same data. Note that the value of m =5 at a lag of -4 for all λ. The confidence interval shown is that for ρ, as this is where the values of m have true predictive value.

We vary the relative phase delay between the two series by lagging φ_y relative to φ_x by a phase lag, Δ:

$$\rho = \frac{1}{N}\sqrt{\left(\sum_{t=1}^{N}\cos(\phi_t - \theta_t)\right)^2 + \left(\sum_{t=1}^{N}\sin(\phi_t - \theta_t)\right)^2} \tag{4}$$

Significance testing of ρ is done by Monte Carlo methods against 1000 realizations of a red noise background (Grinsted et al., 2004), and the results can be visualized in a two-dimensional plot of ρ in λ-Δ space analogous to the wavelet frequency-time space plot. As a further refinement in the utility of such a plot we find it useful to contour the strength of linear regression of the wavelet filtered time series as a function of λ and Δ, so that the color scale bar corresponds to the value of *m* in the equation of $W_Y(\lambda,t+\Delta) = m\ W_X(\lambda,t)$. The phase relationship over the range multi-year to decadal periods was examined by filtering both time series with a Paul wavelet with λ between the Nyquist frequency and 40 years with six λ per octave of scale.

To illustrate the method, we show an example (Figure 25.1) using series where the *X* is red noise with a first order regressive coefficient of 0.8, mean of zero and unit variance, and series *Y* and equal to 5*X* plus white noise (zero mean, unit variance). *X* is then lagged by 4 time units relative to the *Y*, so that it in our sense it *Y* leads and hence is causative of *X*. Simple regression analysis yields a linear best fit the data: Y = 1.9X + 0.61, which fails to capture any important linkage between the two noisy series. However, applying the phase coherence test immediately yields a region of significant coherence, and at a lag of -4 units an in-phase relationship at all filtering periods exists. Finally the bottom panel in Fig. 25.1 shows the magnitude of the lagged regression fit, with an obvious peak value of 5 over all filtering periods at a lag of -4 units.

3 Results

We now use the lag-coherence method to investigate phase relationships between solar irradiance and the atmospheric circulation indices represented by the Arctic Oscillation (AO) and Southern Oscillation Index (SOI). Figure 25.2 shows the sensitivity of the AO to TSI_L and TSI_A. It is clear that there is little of the λ-Δ plot where the phase coherence, ρ is significant, except for the long period TSI_L plot, essentially due to trends in the AO and the irradiance data.

However, the majority of the significant area is in the negative Δ part of the plot, implying that TSI_L would be driving solar cycle – which is physically meaningless. The sensitivity values, especially in the small regions of significant ρ at positive Δ in both panels of Fig. 25.2 are almost absent at period longer than 2 years.

In comparison the TSI links with the SOI (Figure 25.3) show no indication of causality – neither physically meaningful values of lag, nor relationships on any scale with irradiance.

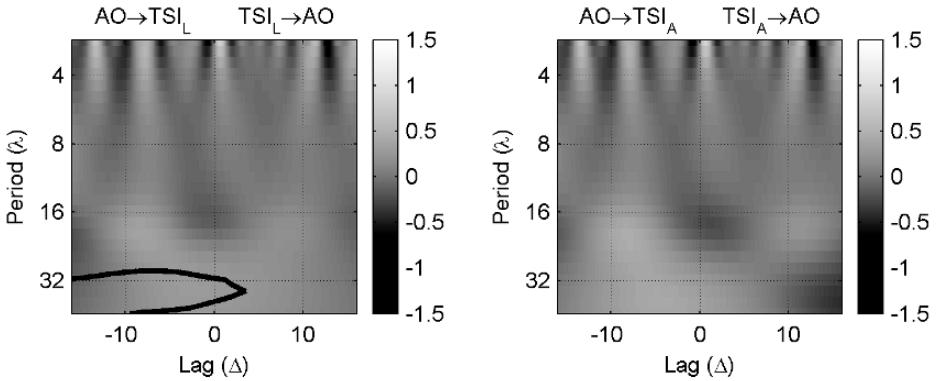


Fig. 25.2. The sensitivity of the monthly Arctic Oscillation index on total solar irradiance based on length (Left panel) and amplitude (Right) of solar cycle. The 95% confidence level of ρ is shown as a thick black contour.

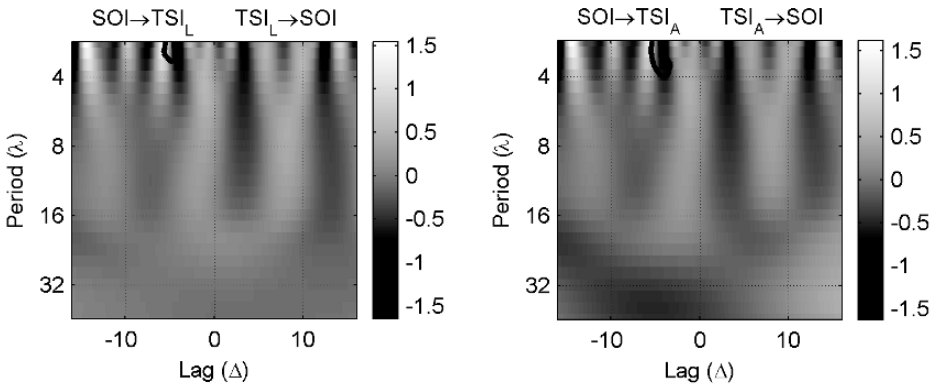


Fig. 25.3. The sensitivity of the monthly Southern Oscillation Index on total solar irradiance based on length (Left panel) and amplitude (Right) of solar cycle. The 95% confidence level of ρ is shown as a thick black contour.

4 Discussion and conclusions

The analysis presented shows clearly that there is no significant causality between solar irradiance and large-scale circulation atmospheric patterns. Thus we extend the analysis of Moore et al. (2006) who showed that simple sunspot numbers had no causal relations with the circulation indices at multi-year to decadal periods. This must cast doubt on the postulated solar UV polar forcing mechanism of climate variability. We have not tested other possible means of solar variability, such as due to solar modulation of cloud intensity via variations in the Earth's magnetic field and

cosmic ray fluxes (Hedfors et al., 2006). While such mechanisms may lead to 11-year periodicities in some climatic indices, the importance of statistically testing their significance against appropriate noise backgrounds is often unappreciated. Of equal importance when causality is alleged, the relative phase of the time series much be tested to verify that hypothesized physical causality is consistent with actual phasing of the time series.

We have shown that a new method of wavelet filtered regression and phase relationship analysis can be used to extract information on lagged responses at specific periods between time series. We show how the significance of the phase relationship can be tested and produce a new type of period-lag plots that shows the regression coefficient between the two time series over a wide range of relative phase lags and frequencies. Thus the specific analysis of a few Paul wavelet periods carried out by Moore et al., (2006) can be extended over the entire range of statistically meaningful periods.

5 Acknowledgments

We thank the Natural Environmental Research Council, the Thule Institute and the Finnish Academy for financial support.

References

Baldwin, M. P., and Dunkerton, T. J. (2001) Stratospheric harbingers of anomalous weather regimes, *Science*, 294, 581–584.

Baldwin, M. P., and Dunkerton, T. J. (1999) Propagation of the Arctic Oscillation from the stratosphere to the troposphere. *J. Geophys. Res.*, **104**, 30,937-30,946.

Baldwin, M. P., and Dunkerton, T. J. (2005) The solar cycle and stratosphere-troposphere dynamical coupling *J Atmos Sol Terr Phys*, 67, 71-82

Barnston, A.G., and Livezey, R.E. (1989) A Closer Look at the Effect of the 11-Year Solar Cycle and the Quasi-biennial Oscillation on Northern Hemisphere 700 mb Height and Extratropical North American Surface Temperature, *J. Clim.* 2, 1295–1313, doi: 10.1175/1520-0442(1989)002

Dijkstra, H.A. and Ghil, M. (2005) Low-frequency variability of the large-scale ocean circulation: A dynamical systems approach, *Rev. Geophys*., 43, RG3002, doi:10.1029/2002RG000122.

Fligge M., and Solanki S.K. (2000) The solar spectral irradiance since 1700, *Geophys. Res. Lett.* 27 (14), 2157.

Grinsted, A., Moore J.C. and Jevrejeva, S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series, *Nonlinear Processes in Geophysics,* 11*,* 561-566.

Hameed, S., and Lee, J. N. (2005), A mechanism for sun-climate connection, *Geophys. Res. Lett.*, *32*, L23817, doi:10.1029/2005GL024393

Hedfors, J., Aldahan, A. Kulan, A. Possnert, G. Karlsson, K.-G.and Vintersved, I. (2006), Clouds and 7Be: Perusing connections between cosmic rays and climate, *J. Geophys. Res*., 111, D02208, doi:10.1029/2005JD005903.

Jevrejeva, S., Moore J.C. and Grinsted, A. (2004). Oceanic and atmospheric transport of multi-year ENSO signatures to the polar regions. *Geophys. Res. Lett.*, 31, L24210, doi:10.1029/2004GL020871

Kodera, K., and Kuroda, Y. (2002) Dynamical response to the solar cycle, *J. Geophys. Res*., 107, 4749, doi:10.1029/ 2002JD002224.

Kuroda, Y., andShibata, K. (2005) Simulation of solar-cycle modulation of the Southern Annular Mode using a chemistry-climate model. *Geophys. Res. Lett.*, **33**, L05703, doi:10.1029/2005GL025095.

Labitzke, K. (2005) On the solar cycle-QBO relationship: a summary *J Atmos Sol Terr Phys*, 67 45-54.

Laut, P. (2003) Solar activity and terrestrial climate: an analysis of some purported correlations *J Atmos Sol Terr Phys* 65 801– 812

Mokhov, I. I. and Smirnov, D. A (2006) El Niño–Southern Oscillation drives North Atlantic Oscillation as revealed with nonlinear techniques from climatic indices *Geophys. Res. Lett*., 33, L03708 10.1029/2005GL024557

Moore, J.C., Grinsted A. and Jevrejeva, S. (2006) Is there evidence for sunspot forcing of climate at multi-year and decadal periods? *Geophys. Res. Lett.* L17705*,* doi:10.1029/2006GL026501

Moron, V., Vautard, R. and Ghil, M. (1998) Trends, interdecadal and interannual oscillations in global sea-surface temperatures. *Clim. Dyn.***14**, 545–569.

Ropelewski, C. F., and Jones, P. D. (1987), An extension of the Tahiti-Darwin Southern Oscillation Index, *Mon. Weather Rev*., 115, 2161– 216

Solanki S.K. and Krivova, N.A. (2003) Can solar variability explain global warming since 1970? *J. Geophys. Res*. 108, 1200, doi:10.1029/2002JA009753.

Thompson, D. W. J. and Wallace, J. M. (1998). The Arctic Oscillation signature in the winter geopotential height and temperature fields. *Geophys. Res. Lett.* **25**, 1297–1300.

Torrence, C. and Compo, G. P. (1998) A practical guide to wavelet analysis, *Bull. Am. Meteorol. Soc.*, 79, 61–78,.

Tsiropoula G. (2003) Signatures of solar activity variability in meteorological parameters *J Atmos Sol Terr Phys*, 65 469-482

26 From Diversity to Volatility: Probability of Daily Precipitation Extremes

Anna K. Panorska[1], Alexander Gershunov[2] and Tomasz J. Kozubowski[1]

[1] Department of Mathematics and Statistics, University of Nevada Reno
[2]Climate Research Division, Scripps Institution of Oceanography

Abstract. A sensible stochastic model is required to correctly estimate the risk associated with daily precipitation extremes. The same requirement holds for studying high-frequency precipitation extremes in the context of climate variability and change. Results derived from probability theory were used to develop an efficient automated scheme to distinguish between heavy and exponential precipitation probability density function (PDF) tails in hundreds of daily station records spanning five decades over the North American continent. These results suggest that, at a vast majority of the stations, daily extreme precipitation probabilities do not decay exponentially, but more closely follow a power law. This means that statistical distributions traditionally used to model daily rainfall (e.g. exponential, Weibull, Gamma, lognormal) generally underestimate the probabilities of extremes. The degree of this distortion, i.e. volatility, depends on regional and seasonal climatic peculiarities. By examining geographical and seasonal patterns in extreme precipitation behavior, the authors show that the degree of volatility is determined regionally by the diversity in precipitation-producing mechanisms, or storm type diversity. Exponential tails are geographically limited to regions where precipitation falls almost exclusively from similar meteorological systems and where light probability tails are observed in all seasons. Topography plays an important role in flattening or fattening PDF tails by limiting the spatial extent of certain systems while orographically altering their precipitation amounts. Results presented here represent the first logical step towards choosing appropriate PDFs at various locations by specifying their regionally relevant family. Heavy tailed models are generally superior to those from the exponential family and can lead to more realistic estimates of extreme event probabilities, return periods, n-year events, and design limits. The correct choice of PDF is essential to safe engineering design, hazard assessment and other applications, as well as for fostering further investigations of hydrologic weather extremes and climate.

1 Introduction

High frequency hydrologic extremes are notoriously volatile and, therefore, difficult to model with traditional probability density functions (PDFs), those with exponentially decreasing probabilities of extreme occurrences. Nevertheless, because they are traditional, exponentially tailed PDFs are typically used to model hydrologic data with the result that observed extremes are left out as improbable events or outliers. This situation leads to traditionally inadequate treatment of daily precipitation extremes in climate research as well as to underestimation of extreme event probabilities in applications which require estimates of environmental risk. Attempts to determine the correct PDF for daily rainfall amounts have so far been either locally focused or assumed simple models for rainfall production universally with the confusing consequence that we still do not know what environmental factors determine the structure of probability tails for daily precipitation. Contradictory results using inconsistent methodologies applied at various locations yield various precipitation PDFs with either exponential or heavy (power law) tails. In this chapter, we try to solve the classification problem in a mathematically, methodologically and spatially consistent way by letting daily precipitation data at hundreds of stations across North America show us the nature of its tails.

The correct choice of PDF is not trivial. For rainfall, a hodgepodge of available theoretical PDFs with either exponential or heavy tails has so far complicated this choice. The data-based approaches to choosing correct PDFs for daily rainfall mostly utilized so far involved fitting specific PDFs to data at limited locations (e.g. Katz et al. 2002). Because, such data-based approaches are methodologically inefficient for choosing correct PDFs at a multitude of available observing stations with daily rainfall, geographic variability in extreme rainfall behavior has not been examined. Alternatively, attempts to determine a "universal" PDF for modeling daily rainfall everywhere have also been made. Most recently, Wilson and Toumi (2005) assumed a simplified model for rainfall production without regard to storm type and then mathematically derived a stretched (i.e. double or extended) exponential PDF as the universal model for extreme rainfall events. In this type of work, data is used as a means to validate the simplified theoretical result, but not to achieve it. Taken in their entirety, previous results do not agree on the existence of a universal PDF for daily rainfall. Nor, do previous results agree on which family of PDFs (i.e. exponentially- or heavy-tailed) best describes rainfall extremes at most locations. We do not know the spatial patterns of applicability of different PDFs for extreme rainfall modeling and, needless to say, we do not have an understanding of the physical processes that may be responsible for any such possible spatial variability or dominance.

In this chapter, we develop and apply an efficient methodology to determine whether exponentially- or heavy-tailed *families* of PDFs best describe the data at hundreds of stations across North America. This data-based *family* classification

problem is, in our thinking, the first step towards understanding hydrologic weather extremes in a climatic context. Fitting specific PDFs at specific locations should be greatly facilitated once such classification is done. Moreover, even a simple visual analysis of the classification result applied to hundreds of stations across the North American continent, should illuminate the physical causes for any possible geographic patterns in the dominance of one family over another.

Statisticians have tried to fit heavy tailed models to hydrologic data. Heavy tailed behavior has been documented in precipitation and streamflow measured at specific locations or regions for time series of extreme values, e.g. annual peak record streamflow (Smith 1989, Katz et al. 2002) and precipitation (Katz et al. 2002), as well as threshold exceedances for streamflow (Anderson and Meerschaert 1998). Heavy tails have also been *suggested* as reasonable models for daily precipitation amount at a large number of U.S. stations (Smith, 2001). However, a formal and efficient classification scheme for heavy vs. exponential probability tails has not been developed and, in climate and hydrology research, exponentially tailed PDFs are still typically used to model daily rainfall, including extremes (e.g. Tsonis 1996, Groisman et al. 1999a, Zolina et al. 2004, Wilson and Toumi 2005).

Heavy tailed distributions such as Pareto (Johnson et al. 1994) or the stable laws (Samorodnitsky and Taqqu 1994) arise naturally as approximations to sums of random quantities such as precipitation accumulated over time. Pareto distribution, via the Peak Over Threshold theory, is also a natural model for excesses over high thresholds. Formally, a random variable X is heavy tailed if $P(X > x) = cx^{-\alpha}$, where c and α are positive constants, for large values of x. Alternatively, we can say that heavy tailed distributions have "power" tails. The most important, essentially qualitative, difference between the heavy and exponentially tailed random quantities is in the size of their large percentiles relative to the majority of the data. The heavy tailed distributions have much larger high percentiles than the exponentially tailed laws, which reflects their higher volatility. We aim to quantify volatility of daily precipitation observed at hundreds of stations over North America and to examine the spatial structure of the result. To quantify volatility and to classify station precipitation into exponential or heavy tailed families of distributions, we use a novel approach stemming from a combination of physical reasoning and a careful examination of the connection between data and mathematics.

We approach the heavy vs. exponential tail classification problem in the context of peaks over threshold (POT) methodology and derive a new statistical test based on the theory of maximum likelihood ratios. Theory, presented in Section 2, implies a broader impact for our test in that it classifies entire PDFs of daily precipitation into exponentially- and heavy-tailed families of distributions. In section 3, we find that daily precipitation probability tails are heavy, not exponential, at a great majority of stations. Moreover, a geographic dependence of volatility is discovered, whereby the heaviest tails occur where precipitation is produced by a variety of storm types, while exponential or light tails occur where diversity in precipitation-producing systems is limited. Important orographic and seasonal controls on precipitation

volatility emerge. A practical impact of these results on estimation of extreme event return periods is briefly considered. Section 4 presents a summary of results and discusses some broad implications for applications and climate research.

2 Methodology and Data

We want the data to show us the nature of its tails: are they exponential or heavy? Aiming to do such a comparison, we are immediately confronted by the fact that many choices exist for the specific form of the PDF with exponential and heavy tails. We consider a random variable X to be heavy tailed if $P(X > x) = cx^{-\alpha}$, where c and α are positive constants, for large values of x. Note, that according to this definition, none of the distributions traditionally used to model precipitation (e.g. Gamma, lognormal, Weibull, double, stretched or extended exponential, etc.) are heavy tailed. Examples of heavy tailed distributions include Pareto, Cauchy, Fréchet, and stable laws. To resolve the problem of distributional multiplicity, i.e. to avoid having to choose specific PDFs from the exponential and heavy-tailed families, we applied the results of the peak-over-threshold (POT) theory. Since our primary interest is in estimating probabilities of large (extreme) events, the POT method helps focus the search for the reasonable models and provides a rigorous mathematical foundation for the results.

2.1 Mathematical theory and statistical test

The POT method involves examination of the data falling above a threshold. For any data value X and threshold u, $X^{[u]}$ is the *exceedance*, that is the (conditional) value of X given that X exceeds u. We consider *excesses* over the threshold u, i.e. $X^{[u]}$-u. The Balkema – de Haan – Pickands theorem (Balkema and de Haan 1974, Pickands 1975) provides the limiting distribution of excesses. The theorem states that when the threshold (u) increases, the distribution of the excesses ($X^{[u]}$-u) converges to a Generalized Pareto (GP) distribution. Any GP distribution has to be one of the following three kinds: exponential, Pareto or Beta. The importance of this result for practical applications is: *no matter what the specific original distribution of X is, the excess, $X^{[u]}$-u, over threshold u has (approximately) one of only three distributions.* The three distributions of the excesses correspond to the tails of the original distribution of X. If X has exponential tail, then $X^{[u]}$-u will have an approximately exponential distribution. If X has heavy tail, then $X^{[u]}$-u will have an approximately Pareto distribution. Beta distribution has finite support and we do not consider it here. Our problem of finding the correct PDF for describing daily extreme event probabilities reduces, therefore, to classifying excesses at a particular station. We

seek a decision rule, ideally a formal statistical test, for classifying the precipitation excesses over threshold into either exponential or Pareto models.

We approached this problem using ideas from the theory of likelihood ratio tests (Lehmann, 1997). Formally, we test the null hypothesis

H_0: data comes from an exponential distribution,
versus the alternative
H_1: data comes from a Pareto distribution.

The approach is to consider the ratio of the maxima of the likelihoods of the observed sample under the Pareto or exponential (in the numerator) and exponential (in the denominator) models. The logarithm[1] of the likelihood ratio, the L statistic, is:

$$L = \log\left(\frac{\max\left(\sup_{\alpha>0, s>0} L_{Pareto}(\vec{x}\,|\,\alpha, s), \sup_{\sigma>0} L_{\exp}(\vec{x}\,|\,\sigma)\right)}{\sup_{\sigma>0} L_{\exp}(\vec{x}\,|\,\sigma)} \right),$$

where $\vec{x}$ is the observed sample of excesses and $L_{Pareto}(\vec{x}\,|\,\alpha, s)$ and $L_{\exp}(\vec{x}\,|\,\sigma)$ are the likelihood functions of the sample under Pareto and exponential models, respectively. We use a Pareto distribution with the survival function $S(x) = P(X>x) = (1/(1+x/s\alpha))^{\alpha}$ and exponential distribution with the survival function $S(x) = P(X>x) = \exp(-x/\sigma)$. To compute L, both likelihoods are maximized first (via maximum likelihood estimates, MLEs, of the parameters), and then the natural logarithm of their ratio is taken as the likelihood ratio statistic.

To help the reader implement the numerical routines necessary for the computation of L, we include several formulas required for the maximum likelihood procedure. Let $\vec{x} = x_1, \ldots, x_n$ be a sample of size n. The supremum of the natural logarithm of the exponential likelihood for this sample is

$$\sup_{\sigma>0} \log(L_{\exp}(\vec{x}\,|\,\sigma)) = n(-\log(\bar{x}) - 1),$$

where $\bar{x}$ is the sample mean. The natural logarithm of the supremum of the Pareto likelihood is

$$\sup_{\alpha>0, s>0} \log(L_{Pareto}(\vec{x}\,|\,\alpha, s)) = n(\log(\hat{\alpha}) - \log(\hat{s}) - 1 - 1/\hat{\alpha}),$$

[1] All logarithms (log) are natural logarithms.

where $\hat{\alpha}$ and $\hat{s}$ are the MLEs of α and s for the Pareto likelihood. The computation of $\hat{s}$ requires numerical maximization of the function

$$u(t) = -\log\left(\frac{\sum_{i=1}^{n}\log(1+x_i t)}{n}\right) + \log(t) - \frac{1}{n}\sum_{i=1}^{n}\log(1+x_i t),$$

with respect to t > 0. If $\hat{t}$ is the maximum of u(t), then the MLE of s is $\hat{s}=1/\hat{t}$. The MLE of α is

$$\hat{\alpha} = \frac{1}{(1/n)\sum_{i=1}^{n}\log(1+\frac{x_i}{\hat{s}})}.$$

The properties of the test, proofs and more details on the optimization process will be published separately (Kozubowski et al., manuscript in preparation).

The limiting distribution of the test statistic L is a mixed distribution, with an atom at zero and distribution given via the distribution of Y=2L:

$$Y = \begin{cases} 0 & \textit{with probability } 1/2 \\ \chi_1^2 & \textit{with probability } 1/2 \end{cases}$$

Using the formula above, we can compute (1-γ)100 percentiles for Y as follows. Any percentile below the 50th percentile is 0. For $0 \le \gamma \le 0.5$, c = (1-γ)100 percentile for Y if $P(Y \le c) = 1-\gamma$, which is equivalent to $P(\chi_1^2 \le c) = 2(1-\gamma)-1$ and thus c can be found in the χ_1^2 tables. Then, the (1-γ)100 percentile of L is c/2.

The (1-γ)100 percentiles of L provide the critical numbers for our test on the significance level γ. The test is one-sided and we reject the null hypothesis if the computed value of the test statistic exceeds the critical number. We have computed some common percentiles for the distribution of L under the null hypothesis for different sample sizes and for the limiting case. The percentiles for finite sample sizes were computed via Monte Carlo simulation with 10,000 samples of a given size from the exponential distribution (Table 26.1).

Table 26.1. The entries are the (1-γ)100 percentiles of the distribution of L for various sample sizes and the limiting distribution (last row) of L. These are also critical numbers for testing our hypothesis on different significance levels γ.

CRITICAL VALUES OF L

Sample size	Significance level, **γ** [and confidence (1 - γ)*100]			
	0.01 [99%]	**0.02** [98%]	**0.05** [95%]	**0.1** [90%]
10	1.71128	1.1561	0.62701	0.25703
50	2.15057	1.5768	0.89045	0.48852
100	2.23171	1.71583	0.94963	0.55706
500	2.45615	1.85952	1.18044	0.70439
1,000	2.51298	1.92766	1.22475	0.71376
5,000	2.62019	1.97122	1.27738	0.76095
10,000	2.70307	2.0285	1.30714	0.80146
∞	2.70595	2.10895	1.35275	0.82120

2.2 Intuitive insight

On a more intuitive note, in the Pareto case, the $\boldsymbol{\alpha}$ parameter determines the thickness of its tail and is of primary importance. As $\boldsymbol{\alpha}$ decreases to zero, the tail of the Pareto distribution becomes heavier (larger volatility) causing the probabilities of extremes to increase. The scale parameter $\boldsymbol{s}$ is of secondary importance. In the exponential case, $\boldsymbol{\sigma}$ is the scale parameter. Statistic L is scale invariant. Note, that as the tail parameter $\boldsymbol{\alpha}$ increases, the Pareto survival function converges to the exponential one. We included the limiting case, $\boldsymbol{\alpha} = \infty$, in the numerator of L. This convergence of Pareto to exponential adds particular difficulty to differentiating between the two distributions as $\boldsymbol{\alpha}$ increases (Fig. 26.1).

The intuition behind this approach is as follows. If a sample (i.e. excesses over threshold) comes from a Pareto distribution, statistic L is expected to be positive because the maximum Pareto likelihood is likely to be larger than the exponential one. Thus, the ratio will be greater than one ($L > 0$). If the sample comes from the exponential distribution, then the Pareto likelihood function is likely to be maximized by the limiting value of $\boldsymbol{\alpha}$ (i.e. $\boldsymbol{\alpha}=\infty$), corresponding to the exponential distribution so that the ratio of the two likelihoods is 1 and its natural logarithm is 0. The distribution of L for Pareto samples with various values of $\boldsymbol{\alpha}$ and for exponential samples is shown in Fig. 26.1.

While, in theory, L should be zero for every exponential sample and positive for any Pareto sample, L has its own variability and the theoretical results do not hold exactly in practice. The main question when considering a statistic for a classification

problem is: Does the statistic differentiate between the two distributions in question. We assessed the viability of L using simulations because the true theoretical distribution of L is not known. The results presented in the form of boxplots in Fig. 26.1 confirm that L differentiates between the

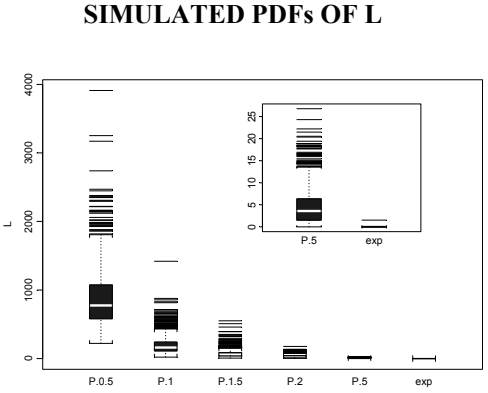


Fig. 26.1. The graph contains 6 boxplots of simulated distributions of L. The first five boxplots were done using 10,000 observations of L from Pareto samples of size 1,000 with α varying from 0.5 (first boxplot) to 5 (second to the last boxplot). The last boxplot corresponds to 10,000 observations of L from exponential samples of size 1,000. The inset blows up the last two boxplots.

Note that the distribution of L for Pareto samples with small α is very different than that for the exponential samples. This provides intuitive evidence that L is able to differentiate between exponential and Pareto distributions where they differ considerably (small α). When the Pareto and exponential distributions become very close to each other (large α), it follows that any probabilities computed from the two (close) models will be very close, so the question of which distribution is better to use becomes mute.

The convergence of Pareto to exponential as α increases is observed as the convergence of the distribution of L for Pareto samples with increasing α to the distribution of L for the exponential samples (Fig. 26.1). Additionally, as α decreases the L statistic becomes "larger on average". This can be visualized by looking at the boxplots and at the values of the median of L that increase as α decreases. This (average) monotonicity of L leads to the conclusion that larger values of L provide, on average, stronger evidence for Pareto distribution. Smaller α's (larger L values, on average) imply heavier tails, which imply larger probabilities of extremes.

We programmed and tested the computation of L from the numerical viewpoint. We also ran several simulation studies and applied the method to the precipitation data with various thresholds. Although threshold choice remains largely intuitive so

far, we obtained fairly robust and meaningful results for daily precipitation with thresholds ranging from the 50th to the 95th percentiles. Below, we present results based on the 75th percentile threshold, which allows a reasonable amount of data for statistical testing in most regions and seasons. In discussing results below, we focus primarily on features generally robust to threshold choice. We do note, however, that optimal threshold choice is recognized as a difficult and still open problem (e.g. Smith 1987, Davidson and Smith 1990, Smith 1994, Gross et al. 1994).

2.3 Data

We have computed L for a subset of over ten thousand available station records of daily rainfall distributed across Canada (Vincent and Gullet 1999), U.S. (Groisman et al. 2004) and Mexico (Miranda 2003) all quality controlled and homogenized at the National Climatic Data Center (NCDC). Five hundred sixty best quality stations were selected to give a reasonable coverage (at least one station within a 70km radius) over North America. The selection used the most stringent quality standards in well-sampled regions and relaxed these standards to include all stations in regions with the poorest coverage (i.e. Alaska, Northern Canada, and parts of Mexico). In mountainous regions where coverage allowed, we always included the highest elevation station along with the best quality station within the 70km radius. At least 80% of the data was required to be present at all stations for the common observational time period: January 1, 1950 to December 31, 2001. In what follows, we compare fits to excesses over the local 75th percentile derived from Pareto to those obtained from the widely used exponential. Figure 26.2 shows the log likelihood ratio statistic (L) at each station.

3 Results

Figure 26.2 is color coded according to the heaviness of the local precipitation tail. This result suggests a spatially coherent dependence of tail type on geographical location with respect to climate. Figure 26.2a presents the absolute magnitude of L. Figure 26.2b more clearly presents the classification result in terms of the statistical confidence with which the null hypothesis (H_o) of exponential tails can be rejected in favor of the heavy-tailed alternative (H_1). The latter result takes sample size into account by interpolating the critical values presented in Table 26.1.

First of all, we see that the vast majority of local tails are strictly speaking non-exponential. In fact, at 81% of the stations, H_o can be rejected with 95% confidence. The degree of departure from the exponential model depends on the peculiarities of regional climatic regimes. Precipitation at stations with a large variety of meteorological influences (i.e. frontal, thunderstorm, tropical cyclone, intense and organized

convection) tends to display highly non-exponential tails. This occurs especially where extreme weather interacts with noteworthy topography (e.g. the Mexican Gulf coast, the front range of the Rockies), the regions of hurricane landfalls (e.g. Gulf and East Coasts of the U.S.), and along Tornado Alley. In the Great Plains and the Midwest, especially the north, heavy precipitation events can appear more extreme relative to typically light winter snow accumulations.

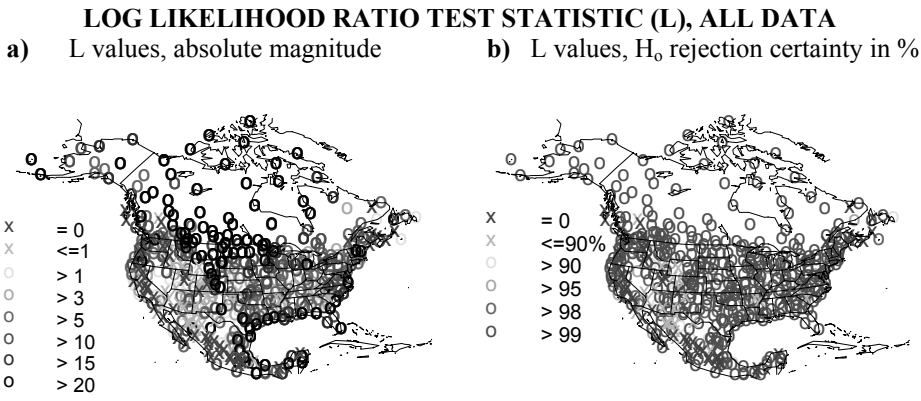


Fig. 26.2. Log likelihood ratio (L) computed for daily excesses over local 75th percentile at each of the 560 stations. **(a)** Values close to zero, L <= 1 (blue and green x's) represent approximately exponential tails, while yellow, red and black circles represent progressively heavier tails. **(b)** Confidence level, (1 - γ)*100, for rejecting the null hypothesis (H_o) of exponential tails. Blue x's represent exponential tails, green x's represent stations at which H_o cannot be rejected with reasonable (90%) confidence. Yellow and progressively redder circles represent stations at which H_o can be rejected with 90, 95, 98 and 99% confidence in favor of the Pareto alternative. H_o can be rejected at 81% of stations with 95% confidence. (A color version of this figure appears between pages 196 and 197).

Exponential behavior occurs where precipitation is reasonably frequent and preferably of one type, i.e. generated by similar type systems. For example, the exclusively convective summer precipitation along the central Mexican plateau as well as isolated regions of almost entirely frontal-system-generated precipitation along the windward side of gently sloping topography (e.g. Sierra Nevada and Rocky Mountains) exhibit exponential tails. Topography appears to play an important role. Notably, precipitation volatility increases abruptly eastward of the Continental Divide at the steep Front Range of the Rockies.

A simple way to categorize precipitation by frontal and convective regimes is to consider winter and summer data separately. Transitional seasons' daily accumulations are interesting as well, as they present preferential mixtures of storm types. Figure 26.3 presents L computed for the December – February (DJF), March – May (MAM), June – August (JJA), and September – October (SON) seasons separately.

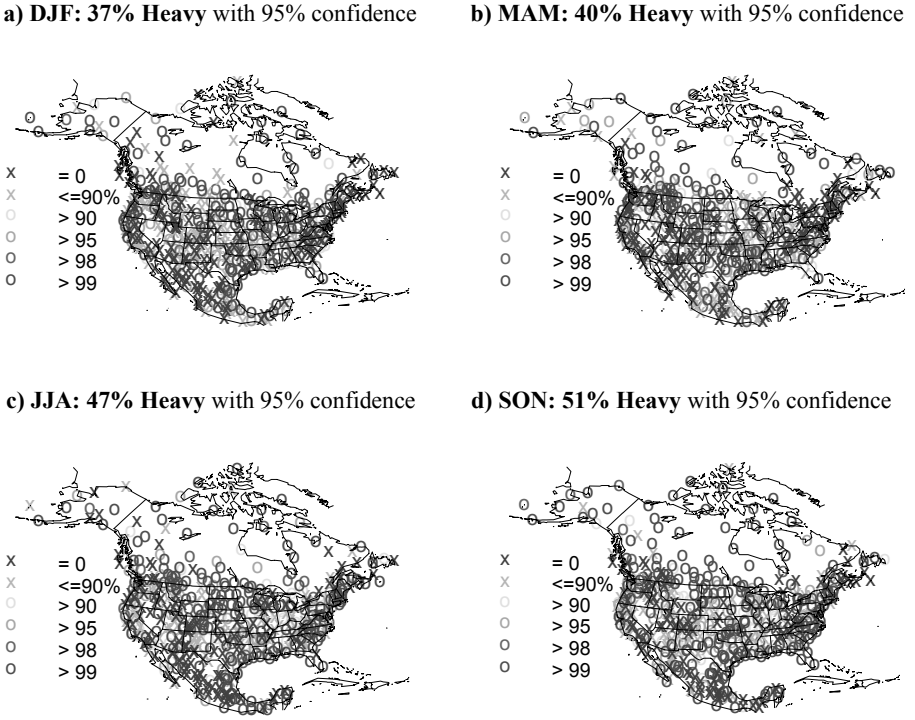


Fig. 26.3. Log likelihood ratio (L) test statistic, as in Fig. 26.2b, but for seasonal data. Plates (a) through (d) show H_o rejection confidence for winter, spring, summer and fall, respectively. Percentage of all stations where exponential tails can be rejected in favor of Pareto with 95% confidence is given in the title for each plate. (A color version of this figure appears between pages 196 and 197).

Precipitation generated by frontal-systems over the mountainous West and the East Coast in winter (Fig. 26.3a) exhibits exponential tails, while the mostly light and moderate cold snow accumulations mixed in with occasional warmer, wetter and heavier events display mostly heavy tails in the northern plains and the Midwestern US. The high Mexican plateau, where it does not rain much in winter, shows mostly exponential tails, while the Mexican part of the Gulf Coast (from southern Texas south to the Isthmus of Tehuantepec), where *nortes* or "northers" bring much of the heaviest winter precipitation exhibits heavy tails. Spring (Fig. 26.3b) mixes convective rainfall with frontal precipitation and brings heavy tails along the US

Gulf Coast and the Deep South. The Continental Divide, visible in all seasons, is best delineated with a strong volatility gradient in spring. Snow accumulations from eastward-moving midlatitude cyclones diminish abruptly east of the Continental Divide, while heavy precipitation occurs here when moist Gulf air is entrapped into the leading edge of south-moving surface anticyclones and slammed against the Front Range.

In summer and fall (Fig. 26.3c,d), heavy tails outline the entire Gulf and East Coasts, where convective rainfall is mixed with that generated by land-falling tropical storms (plus frontal precipitation in fall). The southwestern monsoon fattens precipitation tails in northwestern Mexico and Arizona in summer, while occasional wayward eastern Pacific tropical storms extend heavy tails into Baja California Sur, neighboring Sinaloa, coastal Sonora, and possibly as far north as southern California State in the fall. Along the west coast, the heaviest tails occur in the fall, where typically moderate precipitation amounts can turn very heavy in the occasional early wet-season (most recently, fall 2004 was one of those seasons). The Appalachian range is most clearly outlined with a volatility gradient in the fall. The northwest slope is protected from tropical storm intrusions during the hurricane season, while these storms occasionally reinforce typically lighter precipitation from the more common local and frontal systems southeastward of the range. In general, orography can produce volatility gradients by confining heavy precipitation from specific storm systems to one side of the mountain range, while orographically enhancing their precipitation efficiency. Orography affects storm-type diversity and this effect is typically seasonal.

Generally, L computed for all precipitation data (Fig. 26.2) is an aggregation of seasonal values, reflecting the volatility of precipitation in different seasons (Fig. 26.3). All seasons' precipitation tails are more likely to be heavy at any given station and are heavier than those observed in any one season. For all-season precipitation (Fig. 26.2a), the heaviest tails occur in regions where tails tend to classify heavy in all seasons. These regions are the Gulf Coast, Midwest and the northern plains, and the steepest part of the Front Range of the Rocky Mountains. Exponential tails exist only in regions where precipitation tails are light in all seasons: the gentle west facing slopes of the high western mountains where precipitation is mostly frontal and orographically enhanced, the high plateau of central Mexico where rainstorms are almost exclusively local convective, and at the Canadian Atlantic and Pacific coasts, around drizzly Nova Scotia and Queen Charlotte Islands. Most other regions display overall heavy tails (Fig. 26.2b). The reader is invited to examine Figs. 26.2 and 26.3 to verify that overall heavy tails (Fig. 26.2b) occur in regions where heavy tails appear in at least one of the four seasons. This observation is absolutely consistent with theory, according to which, heavy tails can result by mixing distributions, as long as at least one of those distributions is heavy-tailed.

To help get a visual, intuitive feel for what the measure L represents with respect to the form of the probability tails and their fit to the data, Fig. 26.4 shows probability plots for stations selected to represent specific precipitation regimes and

to roughly span the range of L. These probability plots show how heavy and extreme events are accounted for by exponential and Pareto models. Figure 26.4 presents graphical pairs, scatter plots of the quantiles of the empirical distribution against the quantiles of a theoretical model: exponential or Pareto. As in the calculation of L (Figs. 26.2 and 26.3) the daily station data used here are excesses over the local 75th percentile of the daily amounts recorded on wet days (i.e. heavy to extreme events). In this type of plot, the best model is the one that scatters along a straight line.

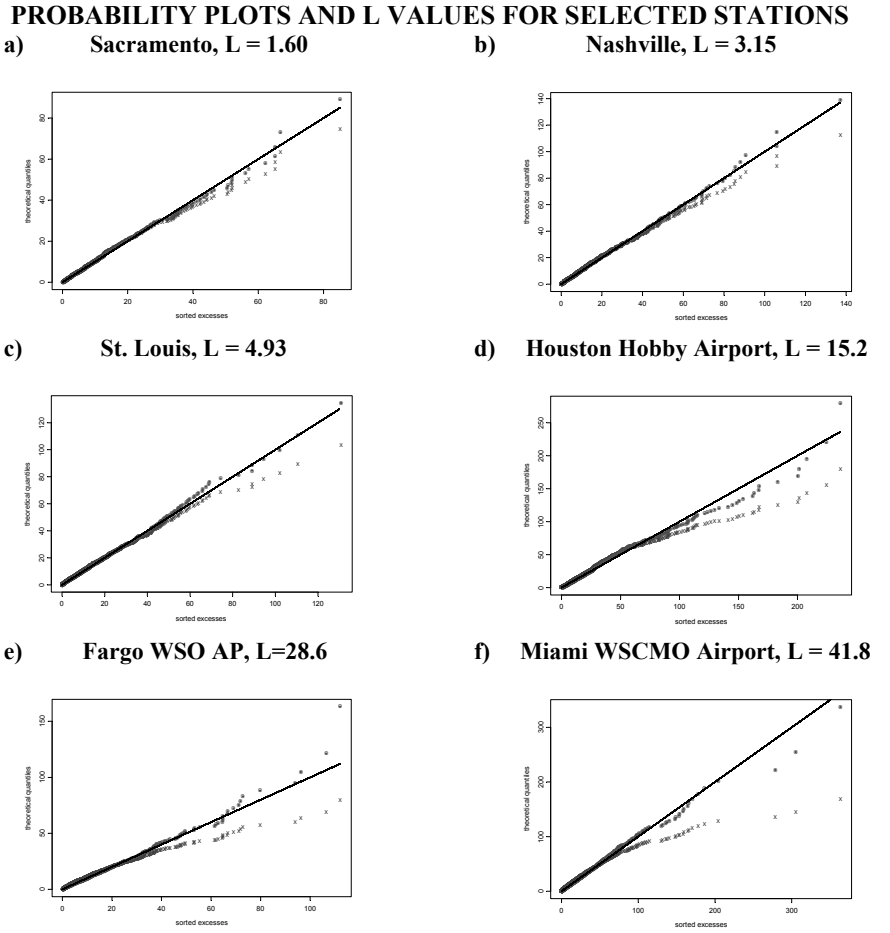


Fig. 26.4. Probability plots for excesses over threshold (75th percentile, see Table 26.1) at selected stations arranged in order of increasing L (see titles for individual plates). Sorted observed excesses displayed in mm along the x-axis are plotted against the corresponding theoretical quantiles derived from the fitted exponential (blue x's) and Pareto (red o's) models.

Table 26.2 presents precipitation statistics at these selected stations for the common observational period 1950–2001: L, probability of precipitation, the 75th percentile of the daily climatology on days with precipitation, maximum recorded daily total, the estimated 100-year event assuming exponential and Pareto tails, and the Pareto probability of exceeding the 100-year event estimated assuming an exponential tail. The thrust of this chapter is on classification. The estimation results presented in the last two columns of Table 26.2, although not rigorous, are shown to illustrate the approximate magnitude that classification can have on natural hazard estimation.

As a rule, where L ≈ 0, both exponential and Pareto tails fit the excesses equally well. We start with an example of a modestly-tailed station, Sacramento FAA AP (40.52°N, 123.82°W, 640m, Fig. 26.4a), located in the Central Valley of California, where rain is produced almost exclusively by frontal systems in the winter half year. Both exponential and Pareto models fit these data reasonably well and L is close to zero. However, the shape of the Pareto tail still differs appreciably from the exponential, especially beyond the observed range. In the 52 years of analyzed data at Sacramento, the 75th percentile of daily total precipitation was 10.7 mm and the maximum was 96 mm. Although both models seem to fit the observed data well, model choice can lead to significant differences in inferences regarding the probabilities of hitherto unobserved extremes. For example, based on the observed

Table 26.2. Precipitation statistics at selected stations for the common observational period 1950 – 2001: L; probability of precipitation (i.e. % of days with recorded precipitation); 75th percentile of daily total on days with precipitation; maximum recorded daily total; the estimated 100-year event assuming exponential and Pareto tails; and the Pareto probability of exceeding the exponential 100-yr event. The last column can be interpreted as the factor by which the 100-yr event estimated assuming exponential tail is more likely to occur assuming Pareto tail. Alternatively, the Pareto return period for an exponential 100-yr event is 100 years divided by the value in the last column at a specific station.

PRECIPITATION STATISTICS AT SELECTED STATIONS

Station	**Log likelihood ratio (L)**	**P[p>0] (%)**	**75th%-ile ($p_{p>0}$) (mm)**	**Max$_{obs}$ (p_{max}) (mm)**	**100yr event** Exp & Pareto **(mm)**	**Pareto P[p>p_{exp}^{100}] (%)**
Sacramento	1.60	16	10.7	96	85 & 99	2.3
Nashville	3.15	26	16	153	127 & 154	3.4
St. Louis	4.93	30	11.2	142	114 & 144	4.1
Houston	15.2	27	16.3	253	195 & 292	6.5
Fargo	28.6	27	5.8	118	85 & 167	12.0
Miami	41.8	36	13.7	377	181 & 346	9.8

record and assuming the exponential (Pareto) model, the estimated 100-year event is 85 (99) mm/day. The Pareto model implies that the 100-year event estimated by the exponential model is 2.3 times as likely to occur, i.e. a 43-year event.

Exponential tails become progressively less adequate as we consider stations with progressively larger L. This is readily visible on the probability plots (Fig. 26.4b-f) as well as by comparing expected values of 100-year events with the observed 52-year station maxima (Table 26.2). Pareto tails perform reasonably well, certainly better by comparison with the exponential for all stations considered. This is not to say that they perform perfectly well, though. At some stations, Pareto tails tend to overestimate the observed probabilities of extremes, as exemplified by extreme values recorded at Fargo, North Dakota (Fig. 26.4e). At others, they may somewhat underestimate probabilities of the largest extremes, e.g. Miami (Fig. 26.4f). It is clear, however, that the heavy tail model, as a rule, outperforms the exponential.

Heavy and extreme daily precipitation at a vast majority of the stations is more in line with the Pareto model. The comparison of 100-year events estimated using the exponential and Pareto models with the observed 52-year station maxima (Table 26.2) reinforces the choice of heavy over exponentially-tailed models to describe probabilities of extreme precipitation events. Assuming, as our results suggest, that the Pareto model is correct in most cases, the choice of an exponentially-tailed PDF may lead to a severe underestimation of extreme precipitation risk. In Miami, for example, the 100-year event estimated by the exponential model corresponds to a 10.2-year event as estimated by Pareto, i.e. is 9.8 times more likely to occur.

4 Summary and Conclusions

The statistical methodology adapted here allows us to simply evaluate whether the tails of an empirical distribution are exponential or heavy, regardless of the specific form of the PDF from the exponential or heavy tailed families of distributions. So, it is not necessary to test whether the data is or is not exponential, Gamma, Gumbel, Weibull, Double (or sometimes “stretched” or “extended”) exponential or any other exponentially-tailed PDF commonly used to model daily rainfall. These are accepted or rejected under the broad umbrella of the exponentially tailed family of distributions. Although theoretical questions related to the choice of threshold remain to be solved, the Pareto/exponential likelihood ratio for peaks over threshold has proven a useful and efficient tool in diagnosing tail behavior and distinguishing between exponentially- and heavy-tailed families of PDFs.

Results reported here confirm that most of the precipitation exceedance probabilities follow the Pareto law and therefore, most precipitation PDF tails follow a power law (i.e. are heavy), not exponential. Threshold exceedances at stations that may have approximately exponential tails can also be thought of as limiting, “special

cases" of the Pareto model, i.e. Pareto with a very large tail parameter, α. As α increases, the Pareto distribution gets close to the exponential and the L statistics may not be sensitive enough for high values of α. But such cases appear to be rare in North America and, in practice, the heavy tail is clearly superior when all seasons' data is considered. The divergence from exponential is a matter of degrees. The most closely exponential tails are found at stations where similar systems and consistent processes produce virtually all of the precipitation. How far from exponential the tails are depends on location and seasonality. Places that receive precipitation from a variety of systems throughout the year tend to exhibit severely non-exponential precipitation tails. Diversity is the mother of volatility.

An exercise in limiting precipitation diversity, by considering various seasons' daily data separately, indicates that daily precipitation tails are heavier when all seasons' data is included compared to individual seasons, further illustrating the idea that the diversity of precipitation-producing systems is essentially responsible for precipitation volatility. This is also clear from examining the geographical patterns of volatility in all, as well as in individual, seasons. Seasonal analyses also indicate that overall heavy tails occur only where at least one season's data classifies as heavy-tailed. This observation is consistent with theory, according to which, a mixture of distributions can yield a heavy tail, only if at least one of the distributions in the mix is heavy-tailed.

Severe diversity in topography also appears capable to fatten or flatten precipitation tails. Notably heavy precipitation tails observed along the precipitous Mexican part of the Gulf Coast in winter, heavy (light) tails observed to the southeast (northwest) of the Appalachian Range during hurricane season, and heavy (light) tails observed east (west) of the Great Continental Divide, all testify to the additional effects of topography on precipitation volatility. Orographic gradients in volatility can be explained by orographically imposed limits on the diversity of precipitation producing systems. A mountain range can enhance or reduce the precipitation produced by specific types of storms while totally or partially limiting the spatial range of some precipitating disturbances to one side of the mountain.

This work is meant as a first step to relate hydrologic weather extremes to climate. It is based on a physically and mathematically consistent attempt to choose a correct model for describing the probability of extreme precipitation events directly from the stochastic properties of the data. Our empirical results strongly suggest the superiority of heavy tails to the traditional exponential. Our approach, results, and their implications, particularly that the fundamental shape of the precipitation probability tail is geographically specific resulting from multiplicity of precipitation producing mechanisms, differ substantially with other work on this subject, most recently and notably by Wilson and Toumi (2005). We believe that the spatial structure and coherence of our results and their solid grounding in climate and weather patterns over the North American continent demonstrate the benefits of our simple approach based on letting the data show us the nature of its tails. Limiting the distributional possibilities based on the stochastic properties of the data is a positive

step forward because the scientific potential and practical implications of classifying precipitation probability tails into heavy- or exponentially-tailed families are considerable.

The real prospects of climatic change emphasize the need for mathematical models of extremes consistent with reality. The urgency of this problem is amplified by the prospect that global hydrologic change may disproportionately manifest itself in increased frequency of extreme precipitation. This view is supported by theoretical reasoning (Allen and Ingram 2002; Trenberth et al. 2003; Karl and Trenberth 2003) and climate pojections (e.g., Zwiers and Kharin 1998; Kharin and Zwiers 2000, Semenov and Bengtsson 2002; Hegerl et al. 2004; Wehner 2004, Groisman et al. 2005) as well as empirical evidence for increasing trends in the frequency of extreme daily precipitation worldwide (e.g., Tsonis 1996, Easterling et al. 2000; Folland and Karl 2001; Groisman et al. 1999a, 2004, 2005). It is difficult, however, to define and account for high frequency precipitation extremes without a reasonable probability distribution model. The choice between heavy-tailed and exponentially-tailed models is of a qualitative nature. The heavy tailed distributions have much larger high percentiles *relative* to the rest of the data values than the exponentially tailed ones. That implies that in places where heavy tailed models are appropriate, and especially if observed and modeled trends continue, future large events may be much larger than those observed to date. We need to be prepared for such possibilities. The exponentially-tailed models of precipitation will not be able to predict very large (relative to the observed data) events, because their mathematical properties do not allow for such extremes. In a future study, we plan to apply our methodology to quantify precipitation volatility in temporally evolving observational and dynamical modeling contexts.

Wherever stochastic modeling of regional precipitation extremes is of interest, a classification of local precipitation into exponential or heavy-tailed families of PDFs is the logical first step towards choosing a specific PDF to represent local precipitation. The classification exercise presented here can be easily adapted to observations at sub-daily temporal scales, wherever available, to better focus on precipitation intensity. Daily accumulations do not represent precipitation intensity well, especially for short-lived local convective systems, and/or fast-moving disturbances. Dynamical atmospheric models are notoriously deficient in their simulation of precipitation intensity, frequency, and, thus, PDFs. Our methodology can be used as a first step to determine the extent and possible causes of dynamical model shortcomings as well as to develop statistical correction schemes for global and regional dynamical model outputs. This methodology can also be used in studies of other hydrologic or broader environmental extremes resulting from volatile processes, e.g. streamflow and wildfire severity. Wherever coherent spatial patterns in volatility are found, they can guide the estimation/interpolation of environmental risk for data-poor locations. The future is bright with interesting problems.

Acknowledgments

This work was supported through NSF grant ATM-0236898 “Modeling, variability and predictability of North American hydrologic extremes”. Funding was also provided by the California Climate Change Center, sponsored by the California Energy Commission's Public Interest Energy Research Program and by the NOAA Office of Global Programs, under the California Applications Program. We have enjoyed and benefited from interesting discussions with Dan Cayan, Mike Dettinger, Kelly Redmond and Tereza Cavazos and we gratefully acknowledge their input. We thank D. Rominger for programming support. Results of this study contribute to the NATO Science for Peace project (SFP 981044) “Extreme precipitation events: their origins, predictability and impacts”.

References

Allen, M.R. and Ingram, W.J. (2002) Constraints on future changes in climate and the hydrological cycle. Nature 419, 224-232.

Anderson, P.L., and Meerschaert, M.M. (1998) Modeling river flows with heavy tails. *Water Resources Research*, **34**, 2271-2280.

Balkema, A.A., and de Haan, L. (1974) Residual life time at great age. Annals of Probability 2, 792-804.

Bradley, B. O., and Taqqu, M. (2003) Financial risk and heavy tails. In: S.T. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance*. Elsevier Science, Amsterdam, pp. 35-103.

Davidson A.C. and Smith, R.L. (1990) Models for exceedances over high thresholds. J. Roy. Statist. Soc. Ser. B 52, 393-442.

Easterling, D.R., Evans, J.L., Groisman, P.Ya., Karl, T.R., Kunkel, K.E. and Ambenje, P. (2000) Observed variability and trends in extreme climate events: A brief review. Bull. Amer. Meteorol. Soc. 81, 417-425.

Falk, M. (1995) On testing the extreme value index via the pot-method. Annals of Statistics 23, 2013-2035.

Folland, C.K. and Karl, T.R. (2001) Observed climate variability and change. In: J.T. Houghton, et al., (Eds.), *Climate Change 2001: The Scientific Basis. Contribution of Working Group 1 to the Third IPCC Scientific Assessment.* Cambridge University Press, Cambridge and New York, pp. 99-181.

Groisman, P.Ya., Karl, T.R., Easterling, D.R., Knight, R.W., Jamason, P.F., Hennessy, K.J., Suppiah, R., Page, C.M., Wibig, J., Fortuniak, K., Razuvaev, V.N., Douglas, A., Førland, E. and Zhai, P-M. (1999a) Changes in the probability of heavy precipitation: Important indicators of climatic change. Climatic Change 42, 243-283.

Groisman P.Ya., Knight, R.W., Karl, T.R., Easterling, D.R., Sun, B. and Lawrimore, J.M. (2004) Contemporary changes of the hydrological cycle over the contiguous United States: Trends derived from in-situ observations. J. Hydrometeor. 5, 64-85.

Groisman, P. Ya., Knight, R.W., Easterling, D.R., Karl, T.R. and Hegerl, G.C. (2005) Trends in intense precipitation in the climate record. J. Climate 18, 1326-1350.

Gross, J., Heckert, A., Lechner, J. and Simiu, E. (1994) Novel extreme value estimation procedures: Application to extreme wind data. In: J. Galambos et al. (Eds.), *Extreme Value Theory and Applications*, Kluwer Academic Publishers, Dordrecht, pp.225-284.

Hegerl, G.C., Zwiers, F.W., Stott, P.A. and Kharin, V.V. (2004) Detectability of anthropogenic changes in temperature and precipitation extremes. J. Climate 17, 3683-3700.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994) *Continuous Univariate Distributions*, 2nd edition, Wiley, New York.

Karl, T.R. and Trenberth, K.E. (2003) Modern global climate change. Science 302, 1719-1723.

Katz, R.W., Parlange, M.B. and Naveau, P. (2002) Statistics of extremes in hydrology. Advances in Water Resources 25, 1287-1304.

Kharin, V.V., and Zwiers, F.W. (2000) Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere-ocean GCM. J. Climate 13, 3760-3788.

Kozubowski, T.J., Panorska, A.P., Qeadan, F. and Gershunov, A. (2007) Testing exponentiality versus Pareto distribution via likelihood ratio. To be submitted. Preprint available from authors.

Lehmann, E.L. (1997) *Testing Statistical Hypotheses*, 2nd edition, Springer, New York.

Miranda S. (2003) Actualizacion de la base de datos ERIC II. Final report of the project TH-0226, IMTA internal reports.

NCDC (2003) Data documentation for data set 3200 (DSI-3200): Surface land daily cooperative summary of the day. National Climatic Data Center, Asheville, NC. [www.ncdc.noaa.gov/pub/data/documentlibrary/tddoc/td3200.pdf].

Pickands, J. (1975) Statistical inference using extreme value order statistics. Annals of Statistics 3, 119-131.

Samorodnitsky, G. and Taqqu, M. (1994) *Stable non-Gaussian Random Processes.* Chapman & Hall, New York.

Semenov V.A. and Bengtsson, L. (2002) Secular trends in daily precipitation characteristics: Greenhouse gas simulation with a coupled AOGCM. Climate Dynamics 19, 123-140.

Smith, J.A. (1989) Regional flood frequency analysis using extreme order statistics of the annual peak record. Water Resources Research 25, 311-317.

Smith, R.L. (1987) Estimating tails of probability distributions. Annals of Statistics 15, 1174-1207.

Smith, R.L. (1994) Multivariate threshold methods. In: J. Galambos et al. (Eds.), *Extreme Value Theory and Applications*, Kluwer Academic Publishers, Dordrecht, pp. 225-284.

Smith, R.L. (2001) Extreme value statistics in meteorology and the environment. *Environmental Statistics*, http://www.stat.unc.edu/postscript/rs/envnotes.pdf, Chpt. 8, 300-357.

Trenberth, K.E., Dai, A., Rasmussen, R.M. and Parsons, D.B. (2003) The changing character of precipitation. Bull. Amer. Meteorol. Soc. 84, 1205-1217.

Tsonis A.A. (1996) Widespread increases in low-frequency variability of precipitation over the past century. Nature 382, 700-702.

van Montfort, M.A.J. and Witter, J.V. (1985) Testing exponentiality against generalized Pareto distribution. Journal of Hydrology 78, 305-315.

Vincent, L.A., and Gullett, D.W. (1999) Canadian historical and homogeneous temperature datasets for climate change analyses. Int. J. of Climatology 19, 1375-1388.

Wehner, M.F. (2004) Predicted twenty-first-century changes in seasonal extreme precipitation events in the Parallel Climate Model. J. Climate 17, 4281-4290.

Wilson P.S. and Toumi, R. (2005) A fundamental probability distribution for heavy rainfall. Geophys. Res. Lett. 32, L14812, doi:10.1029/2005GL022465.

Zolina, O., Kapala, A., Simmer, C. and Gulev, S.K. (2004) Analysis of extreme precipitation over Europe from different reanalyses: A comparative assessment. Global and Planetary Change 44, 129-161.

Zwiers, F.W. and Kharin, V.V. (1998) Changes in the extremes of the climate simulated by CCC GCM2 under CO2 doubling. J. Climate 11, 2200-2222.

27 Stochastic Linear Models of Nonlinear Geosystems

Cécile Penland[1]

[1] NOAA/ESRL/Physical Sciences Division, Cecile.Penland@noaa.gov

Abstract. When nonlinearities are strong enough, they can often be treated in terms of linear stochastic differential equations. We discuss the conditions under which this can be done, and illustrate the linear approximation with several examples having varying success. In particular, we consider the low-dimensional chaotic Lorenz system, a linear process driven by the chaotic Lorenz system, and the real-world application to tropical sea surface temperatures.

1 Introduction: Random Thoughts on Nonlinearity

It has been long known that dynamic systems of different spatial and temporal scales are often coupled. That is, linear low-pass or linear high-pass filters, while often useful, are also limited in applicability. Fortunately, if the behavior of the long timescales is of interest and the detailed behavior of the short ones is less so, it is often possible to estimate the effect of the fast processes on the slower ones through stochastic parameterization. The procedure for developing the stochastic parameterization of rapidly varying processes is far from arbitrary and requires some statistical knowledge of the "uninteresting" fast dynamical component.

The very use of the relative terms "fast" and "slow" indicates a separation of time scales. The limit theorems (Khasminskii 1966; Papanicolaou and Kohler 1974; Majda et al. 1999) describing asymptotic stochastic behavior quantify time scale separation by introducing a smallness parameter ε into the equations of motion. Powers of ε are written as coefficients of individual terms in the equations and indicate the rapidity with which these terms' particular autocorrelation functions decrease with respect to each other. Although ε is a smallness parameter, it is not an importance parameter; terms multiplied by higher powers of ε are not only retained but represent the "interesting" deterministic dynamics. Said another way, we divide the dynamical evolution of a system into large effects that don't last very long and smaller effects that behave coherently. The short-lasting effects are quantitatively treated as Gaussian white noise, possibly modulated by the coherent processes, on a coarse-grained timescale short enough to resolve the slow dynamics but too long to resolve the details of the fast dynamics. The validity of treating rapidly varying components as Gaussian white noise requires that the "noise" be the combined effects of a variety of weakly interacting physical phenomena. This is the reason for coarse-graining the dynamical equations; the unit of time is large enough that the integrated effects of unresolved dynamics can be approximated as a Gaussian stochastic variable (the Central Limit Theorem).

Naturally, the most interesting cases are generally those in which a clear timescale separation is not possible. There are several ways around this problem, although the

mathematical underpinnings are not always as rigorous as when the separation is evident. For example, one might model the stochastic perturbations as red noise with an appropriate spectrum rather than as white noise (e.g., Horsthemke and Léfèver 1984; Newman et al. 1997). This procedure works well as long as one has knowledge of what that spectrum is, and if one is working on a "forward" problem that is *a priori* completely defined. The so-called "inverse problems," where one assumes only the basic form of the dynamical equations and estimates the quantitative description from data, are more difficult.

The difficulty in inverse problems lies in the non-uniqueness of the estimated results, and this difficulty is an issue whether or not there exists a clear timescale separation. Even when it is known that a system is governed by a matrix *linear* equation, with a coefficient of Gaussian white stochastic forcing varying *linearly* with the state vector, one can use sample moments to deduce the matrix coefficient of the noise only up to an arbitrary orthogonal matrix factor. Further, since this article deals with systems where the noise is temporally continuous and only *approximately* white, there is a noise-induced modification (the so-called "Ito correction" or "noise-induced drift," depending on the community) to the systematic linear feedback involving this unobservable orthogonal matrix (Wong and Zakai 1965; Khasminskii 1966). If one is interested only in the predictability of the system, or in covariance statistics of the noise, the arbitrary matrix factor is irrelevant. However, if one is interested in using the results of the inverse model for process studies investigating the detailed interaction between stochastic and deterministic dynamics, the arbitrariness can be problematic. In this study, we shall use inverse modeling to diagnose covariance statistics of a low-order deterministic chaotic process acting as additive forcing to a linear system, thus avoiding the issue of non-unique results.

The idea that low-order chaos can act as stochastic forcing is not new. Jarzynski (1995) has shown that fast deterministic chaos coupled to a slow Brownian particle drives the slow degrees of freedom toward a state of statistical equilibrium with the fast degrees. This result is general, and was achieved by analyzing the Liouville equation for the combined fast-slow system to investigate the probability distribution of the slow part. In other work, Rödenbeck et al (2001) considered a 9-dimensional model introduced by Lorenz (1995), coupled to a faster 8-dimensional system. The important restriction in these studies was the well-separated timescales of the fast and slow dynamics. And so, since the issue of whether driving chaos is of low or high dimensionality is not really important to the validity of a stochastic approximation, we are back to the issue of how far we can go with such an approximation when the timescales are not well-separated.

Why would one want to do such a thing? Why not just consider the nonlinear system as a whole? The problem, of course, is in the word "just;" most of the time we do not know the equations for the full nonlinear system. We may know some of them, but we should not expect a good description of reality by a geophysical model confined to those equations that are well understood, when the effect of unknown processes might be as large as that of the known ones. We can replace the unknown dynamics with detailed conceptual models based on data, with the added advantage of being able to analyze these "toy" models in detail. This is an excellent approach as

long as one is able to justify quantitatively the rejection of competing possibilities, but bears the danger of legitimizing unjustified assumptions, depending on how ad hoc the toy model is. A model ought not be accepted on the basis of a test that can pass mutually exclusive dynamics. For example, prediction error alone is not always a good criterion for evaluating competing physical models; for more than a thousand years before Copernicus, Kepler and Galileo appeared on the scene, Aristarchus' (4th century BCE) heliocentric model of the solar system was routinely rejected in favor of Ptolemy's (2nd century CE) earth-centered model because Ptolemy's model predicted planetary positions more accurately.

If there exists a plethora of detailed deterministic models having different dynamical properties but similar resemblances to observations (this is the case with El Niño, for example), it is often useful to step back to the most general description of that system's dynamics and use the results of the Central Limit Theorem (CLT) to devise a rigorous approximation, including an estimation of how accurate the approximation is on realistic space and time scales. That is, "rigorous" does not always mean "accurate," or vice versa; it does mean that the conditions for application of the CLT (see discussion below) are met.

In the following, we shall explore the completely rigorous, but sometimes inaccurate, approximation that a nonlinear system can be described as a *linear* process augmented with stochastic terms resulting from rapidly varying nonlinearities. It must be emphasized that we are not throwing away the nonlinearities; we retain them in an approximate form. This approximation cannot be made arbitrarily, and so we review its methodology in some detail. In particular, we discuss a test for the validity of the linear approximation that is difficult to pass. This test is subject to "Type 1 errors," often rejecting the linear hypothesis when it is true. Most statistical tests are subject to either Type 1 errors (rejecting the hypothesis when it is true) or Type 2 errors (accepting the hypothesis when it is false). While neither type of error is desirable, Type 2 errors are worse. They are also the errors likely to be made by those who look for evidence in favor of theories rather than for evidence against them.

After considering some theoretical aspects of the analysis, we use the familiar chaotic Lorenz (1963) system to show an example of a rigorous but useless application of the approximation. After that, we show an example where the approximation is somewhat problematic but is still extremely useful in diagnosing many dynamical properties of the system. The next stage of the article applies the approximation to a real geophysical system. Finally, we discuss the results presented here. By elucidating that the dependence of a dynamical description of natural phenomena in terms of linear and/or nonlinear processes is largely a matter of scale, it is hoped that the commonalities rather than the conflicts between different views of dynamical systems will be emphasized.

2 Theoretical Considerations

2.1 The central limit theorem

The first part of this discussion is taken from Sardeshmukh et al. (2001) and is repeated here mostly verbatim because of its importance to this study. (Let the reader be assured that I am plagiarizing myself.) Consider a dynamical system of equations as follows:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{F}'(\boldsymbol{x},t) + \boldsymbol{G}'(\boldsymbol{x},t) \tag{1}$$

where $\boldsymbol{x}$ is a vector in the N-dimensional Euclidean space $\mathbb{R}^N$ and where $\boldsymbol{F'}(\boldsymbol{x}, t)$ and $\boldsymbol{G'}(\boldsymbol{x}, t)$ are characterized by short and long correlation times, respectively. For our purposes, an alternative description of $\boldsymbol{x}$ in terms of a dimensionless parameter ε is preferable

$$\frac{d\boldsymbol{x}}{dt} = \varepsilon\boldsymbol{F}(\boldsymbol{x},t) + \varepsilon^2\boldsymbol{G}(\boldsymbol{x},t) \quad , \tag{2}$$

where ε^2 can be thought of as a ratio of characteristics timescales of $\boldsymbol{F'}$ and $\boldsymbol{G'}$. It should be noted that the parameter ε is not intended here to be a measure of the relative importance of $\boldsymbol{F}(\boldsymbol{x}, t)$ and $\boldsymbol{G}(\boldsymbol{x}, t)$ but rather of the relative rapidity with which the autocorrelation functions of these terms decay, as will be clear in what follows. The theorem of Papanicolaou and Kohler (1974; PK74 hereafter) describes the conditions under which a singular scaling of time allows Eq.(2) to converge weakly to a stochastic differential equation. This is a dynamical form of the Central Limit Theorem. The conditions require that the fast process $\varepsilon\boldsymbol{F}(\boldsymbol{x}, t)$ be sufficiently variable and that the probability density function (pdf) of any value of $\varepsilon\boldsymbol{F}(\boldsymbol{x}, t)$ becomes independent of any initial conditions as time increases indefinitely, and at a sufficiently rapid rate. Further, $\boldsymbol{F}(\boldsymbol{x}, t)$ is required to be sufficiently smooth with respect to the components of $\boldsymbol{x}$, where "sufficiently rapid" and "sufficiently smooth" are made explicit in PK74. The time coordinate is now scaled

$$s = \varepsilon^2 t \tag{3}$$

and Eq.(2) becomes

$$\frac{d\boldsymbol{x}}{ds} = \frac{1}{\varepsilon}\boldsymbol{F}(\boldsymbol{x}, s/\varepsilon^2) + \boldsymbol{G}(\boldsymbol{x}, s/\varepsilon^2) \tag{4}$$

The proof in PK74 that Eq.(4) converges weakly to a stochastic differential equation is very general, and often difficult to apply in a forward sense. Therefore, for illustrative purposes, we restrict the problem by putting more conditions on $\boldsymbol{F}(\boldsymbol{x},s/\varepsilon^2)$, and by stating that $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{x}$ are all vectors with N elements. Let $\boldsymbol{F}(\boldsymbol{x},s/\varepsilon^2)$ be of the form (see Remark 8 of PK74)

$$F_i(\boldsymbol{x}, s/\varepsilon^2) = \sum_{k=1}^{K} F_i^k(\boldsymbol{x}, s)\eta_k(s/\varepsilon^2), \tag{5}$$

where $\eta_k(s/\varepsilon^2)$ is stationary, centered and bounded. The integrated lagged covariance matrix of $\boldsymbol{\eta}$ has elements

$$C_{km} = \int_0^\infty < \eta_k(t)\eta_m(t+\tau) > d\tau, \qquad k,m = 1,2,...K, \tag{6}$$

where angle brackets denote expectation value. With these restrictions, the theorem by PK74 states that in the limit of long times ($t\to\infty$) and small ε ($\varepsilon \to 0$), taken so that $s = \varepsilon^2 t$ remains fixed, the conditional pdf $p(\boldsymbol{x}, s \mid \boldsymbol{x}_0, s_0)$ for $\boldsymbol{x}$ at time s given an initial condition $\boldsymbol{x}_0(s_0)$ satisfies the "backward Kolmogorov equation," so called because the operator $\mathcal{L}$ describes the dependence of $p(\boldsymbol{x}, s \mid \boldsymbol{x}_0, s_0)$ on the initial conditions:

$$\frac{\partial p(\boldsymbol{x}, s \mid \boldsymbol{x}_0, s_0)}{\partial s_0} = \mathcal{L}(p(\boldsymbol{x}, s \mid \boldsymbol{x}_0, s_0)), \tag{7}$$

where

$$\mathcal{L}(\bullet) = \sum_{ij=1}^{N} a^{ij}(\boldsymbol{x}_0, s_0)\frac{\partial^2}{\partial x_{oi}\partial x_{0j}}(\bullet) + \sum_{j=1}^{N} b^j(\boldsymbol{x}_0, s_0)\frac{\partial}{\partial x_{0j}}(\bullet) \tag{8}$$

and

$$a^{ij}(\boldsymbol{x}, s) = \sum_{km=1}^{K} C_{km}F_i^k(\boldsymbol{x}, s)F_j^m(\boldsymbol{x}, s) \tag{9a}$$

$$b^j(\boldsymbol{x}, s) = \sum_{km=1}^{K} C_{km}\sum_{i=1}^{N} F_i^k(\boldsymbol{x}, s)\frac{\partial F_j^m(\boldsymbol{x}, s)}{\partial x_i} + G_j(\boldsymbol{x}, s). \tag{9b}$$

In this limit, if the formal adjoint $\mathcal{L}^*$ of $\mathcal{L}$ exists, the conditional pdf also satisfies a "forward Kolmogorov equation" in the scaled coordinates:

$$\frac{\partial p(\boldsymbol{x}, s \mid \boldsymbol{x}_0, s_0)}{\partial s} = \mathcal{L}^*(p(\boldsymbol{x}, s \mid \boldsymbol{x}_0, s_0)) \tag{10}$$

Where

$$\mathcal{L}^*(\bullet) = \sum_{ij=1}^{N} \frac{\partial^2}{\partial x_i \partial x_j}(a^{ij}(\boldsymbol{x}, s)\bullet) - \sum_{j=1}^{N} \frac{\partial}{\partial x_j}(b^j(\boldsymbol{x}, s)\bullet) \tag{11}$$

and where the superscript asterisk indicates the formal operator adjoint. Eq. (10) is called a "Fokker-Planck equation" in the scientific literature, and we adopt this terminology hereafter. As stated above, the conditional pdf of $\boldsymbol{x}$ in the scaled

coordinate system obeys Eqs. (7) and (10) in a weak sense. That is, the moments of $\boldsymbol{x}$ can be approximated with an error of order ε by the moments of the solution to the stochastic differential equation,

$$d\boldsymbol{x} = \boldsymbol{G}(\boldsymbol{x}, s)\ ds + \sum_{k,\alpha} \boldsymbol{F}^k(\boldsymbol{x}, s) S_{k\alpha} \bullet dW_\alpha\,, \tag{12}$$

where the symmetric matrix $\boldsymbol{C}$ has been written as the product of two matrices ($\boldsymbol{C} = \boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}$) and has absorbed the factor of 1/2 present in most formulations of the Fokker-Planck equation. The quantity $\boldsymbol{W}$ is a vector of independent Wiener processes and the expression $\bullet d\boldsymbol{W}$ denotes the fact that Eq. (12) is to be interpreted in the sense of Stratonovich. That is, the white noise is an approximation to a continuous system with small but finite decorrelation time. One implication of this result is that stochastic integrals reduce to standard Riemann integrals (Kloeden and Platen 1992). Since we are usually interested in moments of the system rather than sample paths, the stochastic integral property of Stratonovich systems is less important than the dependence of moments, especially the mean, of the limiting process on the form of the stochasticity. Unless $\boldsymbol{F'}$ is independent of $\boldsymbol{x}$, the ensemble mean of $\boldsymbol{x}$ will be different from the deterministic solution for the differential equation with only the slow part $\boldsymbol{G'}$.

In forward problems, the form of $\boldsymbol{\eta}$ usually dictates the form of $\boldsymbol{S}$. Inverse problems are usually defined only up to an estimation of $\boldsymbol{C}$ and, therefore, infinitely many matrices $\boldsymbol{S}$ consistent with $\boldsymbol{C}$ may be constructed by multiplying any estimation of $\boldsymbol{S}$ on the right by an arbitrary orthogonal matrix.

As a final comment on this section, we note that the CLT is posed for "random" processes $\boldsymbol{F}(\boldsymbol{x}, t)$. It should be noted that this definition of "random" is from Kolmogorov (1933) and includes the deterministic processes. The most important condition on $\boldsymbol{F}(\boldsymbol{x}, t)$ is the "mixing condition," which states how rapidly the autocorrelation function of $\boldsymbol{F}(\boldsymbol{x}, t)$ decreases. The mixing condition in PK74 is much weaker than that of Khasminsii (1966) and is fairly easy for most physical processes to pass. Beck and Roepstorff (1987) further explored this issue further in the context of chaotic systems. Thus, we interpret the CLT to mean that there is a timescale on which the moments of a deterministic chaotic process are indistinguishable from those of a system governed by the stochastic differential equation (12). Whether or not this timescale is useful is another matter. For example, the approximation is rigorous but useless for prediction studies if the behavior of a nonlinear system is coherent enough that the timescale for the stochastic approximation is too long to resolve the differences between conditional and marginal probabilities. One valuable use of the CLT is in model evaluation. If nature can be shown to obey the CLT at some timescale, then any realistic model of nature must also obey it at that same timescale. We shall explore this point further below.

2.2 Linear inverse modeling

We now explore the possibility of a natural process obeying a rather extreme version of the CLT, the multivariate, stable linear process driven by stochastic forcing. This approximation is surprisingly useful, particularly when the deterministic feedback matrix is not orthogonal.

Linear Inverse Modeling (LIM) is a statistical method of deriving the best linear description from observations of a dynamical system. The first step of this procedure is also known as Principal Oscillation Pattern (POP) analysis (Hasselmann 1988; Von Storch et al. 1988). Unlike POP analysis, however, LIM attempts to identify the dynamical nature of the system being observed. Inherent in the procedure is a test to see whether or not the procedure is appropriate for the time series, i.e., whether the dynamical system generating the time series is, indeed, linear. We also review LIM's ability to describe the transient growth that can occur when the linear dynamical operator is not orthogonal (e.g., Farrell 1988).

Consider a multivariate linear dynamical system $\boldsymbol{x}$ driven by additive, stationary, Gaussian white noise $\boldsymbol{\xi}$:

$$\frac{d\boldsymbol{x}}{dt} = \mathbf{B}\boldsymbol{x} + \boldsymbol{\xi}\,. \tag{13}$$

In Eq. (13), $\boldsymbol{\xi}$ is related to a vector Brownian motion $\boldsymbol{W}(t)$ as follows:

$$\boldsymbol{\xi}\ dt = \mathbf{S}d\boldsymbol{W}, \tag{14}$$

so that $\mathbf{Q} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$ is a matrix measuring the covariance properties of the stochastic forcing. The analysis that follows can be adapted to several variations of Eq. (13). In particular, we may allow $\mathbf{S}$ to vary periodically in time, or to be a linear function of $\boldsymbol{x}$. We shall discuss these adaptations to the procedure during the exposition.

Moments of the Fokker-Planck equation for the transition pdf $p(\boldsymbol{x},t+\tau\,|\boldsymbol{x},t)$ corresponding to Eq. (13) can be manipulated to yield

$$<\boldsymbol{x}(t+\tau)\boldsymbol{x}^{\mathrm{T}}(t)> = \exp(\mathbf{B}\,\tau)\ <\boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)>, \tag{15}$$

where angle brackets denote ensemble averages. In practice, the contemporaneous and lagged autocovariance matrices are estimated as time averages. If $\mathbf{S}$ is a periodic function rather than a constant matrix, Eq. (15) is true as long as the matrices are averaged over an integer number of periods. The difference between this derivation and derivations for discrete analyses by, for example, Von Storch et al (1988) is that their noise is required to be contemporaneously uncorrelated with the state variable. In a continuous system Eq. (13), where the driving noise is a physical process with a nonzero but negligible correlation time, the system $\boldsymbol{x}(t)$ and $\boldsymbol{\xi}(t)$ are correlated as follows (García et al. 1987):

$$<\boldsymbol{x}(t)\boldsymbol{\xi}^{\mathrm{T}}(t)> + <\boldsymbol{\xi}(t)\boldsymbol{x}^{\mathrm{T}}(t)> = \mathbf{S}\mathbf{S}^{\mathrm{T}} = \mathbf{Q}. \tag{16}$$

Principal Oscillation Pattern (POP) analysis (Hasselmann 1988; Von Storch et al. 1988) estimates the eigenstructure of an operator $\mathbf{A}$ that replaces $\exp(\mathbf{B}\,\tau)$ in Eq. (16). POP analysis is the first step of LIM. However, instead of attempting to identify a

single dominant mode, LIM identifies $\mathbf{A}$ with the Green function $\mathbf{G}(\tau) \equiv \exp(\mathbf{B}\tau)$, with τ chosen to optimize the accuracy of the calculation (Penland and Sardeshmukh 1995*a*). The entire set of normal modes (eigenvectors of $\mathbf{G}(\tau)$ and, hence, $\mathbf{B}$) and adjoints (eigenvectors of $\mathbf{G}^{\mathrm{T}}(\tau)$ and, hence, $\mathbf{B}^{\mathrm{T}}$) is then used to infer the dynamical operator $\mathbf{B}$.

One very useful property of using the stationary continuous system is the Fluctuation-Dissipation relation (FDR: e.g., Penland and Matrosova 1994)

$$\mathbf{B}\langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle + \langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle\mathbf{B}^{\mathrm{T}} + \mathbf{Q} = \mathbf{0}. \tag{17}$$

Equation (18) states that the system's stationary statistics, represented by $\langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle$, are maintained by a balance between the stochastic forcing, represented by $\mathbf{Q}$, and the dissipation effected by the deterministic dynamics represented by $\mathbf{B}$. For periodically varying $\mathbf{S}$, the right side of Eq. (18) is not zero, but is rather the derivative with respect to time of $\langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle$.

Given any initial condition $\boldsymbol{x}(t)$, the most probable prediction $\boldsymbol{x}'(t+\tau)$ of $\boldsymbol{x}(t+\tau)$ is performed simply by multiplying the initial condition by the Green function as follows:

$$\boldsymbol{x}'(t+\tau) = \mathbf{G}(\tau)\boldsymbol{x}(t) \equiv \exp(\mathbf{B}\tau)\,\boldsymbol{x}(t). \tag{18}$$

For a perfect model of the stochastic system (13), the prediction error

$$\boldsymbol{\varepsilon}(t+\tau) \equiv \boldsymbol{x}(t+\tau) - \boldsymbol{x}'(t+\tau) \tag{19}$$

is not zero, but is rather a Gaussian random variable with covariance matrix $\boldsymbol{\sigma}$:

$$\boldsymbol{\sigma}(t+\tau) \equiv \langle \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}(t+\tau)\rangle = \langle \boldsymbol{x}(t+\tau)\boldsymbol{x}^{\mathrm{T}}(t+\tau)\rangle - \mathbf{G}(\tau)\langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle\mathbf{G}^{\mathrm{T}}(\tau). \tag{20}$$

For constant $\mathbf{S}$, of course, $\boldsymbol{x}$ is wide-sense stationary. One now has a test for the validity of Eq. (13). A lag τ_o is chosen and Green functions at other lags are estimated using the following relation:

$$\mathbf{G}(\tau) = [\mathbf{G}(\tau_o)]^{\tau/\tau o}. \tag{21}$$

(Eq. (21) is obtained by applying the Cayley-Hamilton theorem to the spectral decomposition of $\mathbf{G}(\tau)$). If Eq. (13) is valid, the prediction error variance does not depend on τ_o. Penland and Magorian (1993) and Penland and Sardeshmukh (1995*b*, PS95 hereafter) both found this "tau test" to be passed by LIM applied to tropical IndoPacific sea surface temperature anomalies.

There are, unfortunately, linear systems for which the tau test is not valid. If $\mathbf{B}$ is not a constant matrix, then the tau test will fail. The tau test will also fail if the dimensionality of the observed state vector is too small to span the linear vector space in which the dynamics reside (Penland and Ghil 1993). Sometimes, the lag τ_o happens to lie close to the half-period of a normal mode's eigenvalue, thereby precluding the diagnosis of that mode's imaginary part and causing an inaccurate estimation of $\mathbf{G}(\tau_o)$. Yet another source of Type 1 errors (rejecting the linear approximation hypothesis when it is true) is observational error in the data from which the model is estimated. Nevertheless, if the validity of the linear approximation is the question, Type 1 errors of this genre are probably safer in

diagnosing the dynamical behavior of a system than Type 2 errors (accepting the linear approximation hypothesis when it is false) would be.

If $\mathbf{S}d\boldsymbol{W}$ is a linear function of $\boldsymbol{x}$ as follows:

$$(\mathbf{S}d\boldsymbol{W})_i = \sum_j S^{(1)}{}_{ij} x_j \bullet dW^{(1)} + \sum_\beta S^{(2)}_{i\beta} dW_\beta{}^{(2)}, \tag{22}$$

with $\mathbf{S}^{(1)}$ a constant matrix, the above formalism follows similarly with some exceptions. First of all, the matrix Green function $\mathbf{G}(\tau)$ becomes

$$\mathbf{G}(\tau) = \exp\{[\mathbf{B} + (\mathbf{S}^{(1)})^2/2]\,\tau\}. \tag{23}$$

Secondly, it may not be possible to prove that the prediction $\boldsymbol{x}'(t+\tau) = \mathbf{G}(\tau)\boldsymbol{x}(t)$ is the most probable prediction (if the existence of a probability current around the boundaries of the phase space cannot be shown to vanish), but $\boldsymbol{x}'(t+\tau)$ is still the best prediction in the mean square sense. Next, Eq. (16) does not hold. Finally, the FDR is modified to read

$$[\mathbf{B} + (\mathbf{S}^{(1)})^2/2]\langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle + \langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle[\mathbf{B} + (\mathbf{S}^{(1)})^2/2]^{\mathrm{T}} + \ldots$$
$$\ldots + \mathbf{S}^{(1)}\langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle\mathbf{S}^{\mathrm{T}(1)} + \mathbf{S}^{(2)}\mathbf{S}^{\mathrm{T}(2)} = \mathbf{0}. \tag{24}$$

The annoyance of these expressions is that it is generally possible to infer only the combination $[\mathbf{B} + (\mathbf{S}^{(1)})^2/2]$ from data, rather than $\mathbf{B}$ and $\mathbf{S}^{(1)}$ separately. However, the tau test for the linear approximation is still valid. Thus, LIM can distinguish linear dynamics from demonstrably nonlinear dynamics in data, but it cannot distinguish an additive from a multiplicative noise process.

For simplicity of explanation, we now return to the case of constant $\mathbf{S}$. The eigenectors of $\mathbf{B}$, and hence $\mathbf{G}(\tau)$, are generally not orthogonal to each other, in spite of their being known as "normal" modes. When $\mathbf{B}$ is orthogonal and $\mathbf{Q}$ is diagonal, then the normal modes are indeed orthogonal and equivalent to the Empirical Orthogonal Functions (EOFs). However, this is a highly specialized case. Whenever the normal modes are orthogonal, the amplitude of a prediction $\boldsymbol{x}'(t+\tau)$ is always smaller than that of the initial condition. When they are not orthogonal, the non-normality sometimes causes temporary amplitude growth (Farrell 1988; Penland and Sardeshmukh 1995*b*). The condition for this temporary growth is that the initial condition project strongly onto a right singular vector of the operator $\mathbf{G}(\tau)$, i.e., an eigenvector of $\mathbf{G}^{\mathrm{T}}\mathbf{G}(\tau)$, and that this singular vector correspond to an eigenvalue greater than one. In fact, the eigenvalue is the ratio of the predicted amplitude at lead time τ to the initial condition amplitude and is called the "growth factor at lead time τ."

3 Experiments Using the Chaotic Lorenz System

In this section we shall consider the deterministic equations of Lorenz (1963):

$$\frac{dx}{dt} = -a(x - y) \tag{25a}$$

$$\frac{dy}{dt} = rx - xz - y \tag{25b}$$

$$\frac{dz}{dt} = xy - bz \tag{25c}$$

where the choice of parameters $a = 10$, $b = 8/3$ and $r = 28$ give the familiar chaotic behavior. Equations (25) are integrated using a simple Euler scheme with a model timestep $\Delta = 0.001$, and sampled every 20 timesteps, so that the sampling interval Δ_s is 0.02. The first 1000 samples were discarded to ensure that the time series used in the experiments are representative of the attractor. The time series $x(t)$, $y(t)$ and $z(t)$ are shown in Fig. 27.1; the usual projections of the attractor are shown in Fig. 27.2. A time series of 500 000 samples (10 000 model time units) was generated; hereafter we call it the *L63 time series.*

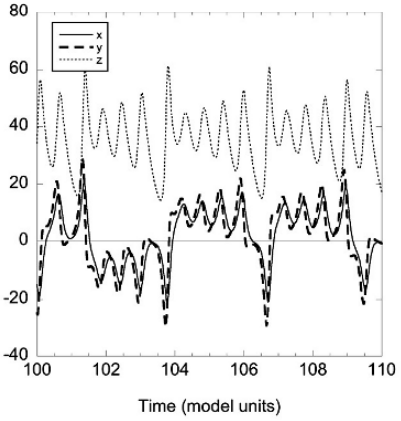


Fig. 27.1. Time series of the chaotic Lorenz (1963) system used in this study. Solid line: x. Dashed line: y. Dotted line: z.

3.1 A linear model of the chaotic Lorenz system

This does not work. It is not expected to work. The reason we tried it is to give credibility to results for systems where the linear approximation *does* work, at least, in comparison.

Let us linearize Equations (25) around $\boldsymbol{l}_0 = (x_0, y_0, z_0) = (0, 0, \langle z \rangle)$ so that the equation for the vector $\boldsymbol{l} = (x, y, \zeta)$, where $\zeta = z - z_0$, can be written

$$\frac{d\boldsymbol{l}}{dt} = \mathbf{L}\boldsymbol{l} + \boldsymbol{\eta} \tag{26a}$$

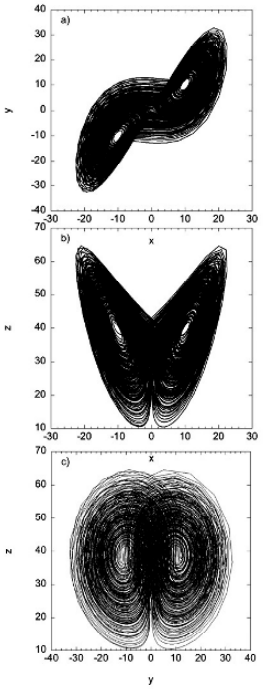


Fig. 27.2. Maps of the chaotic Lorenz attractor generated for this study. *a) y vs. x. b) z vs. x. c) z vs. y.*

with

$$\mathbf{L} = \begin{pmatrix} -a & a & 0 \\ r - z_0 & -1 & 0 \\ 0 & 0 & -b \end{pmatrix} \qquad (26b)$$

and

$$\eta = \begin{pmatrix} 0 \\ -x\varsigma \\ xy - bz_0 \end{pmatrix}. \qquad (26c)$$

To see whether Eq. (26) can be estimated as a stochastic linear process on the sampling time scale, we centered the L63 time series and subjected it to LIM, using values of τ_0 = 0.02, 0.04, 0.06, 0.08, 0.10, 0.12, 0.14, and 0.16. Traces of the

resulting estimates **B'** of **B** are shown in Fig. 27.3. Clearly, **B'** does not pass the tau test. In fact, trace(**B'**) behaves exactly as expected for a nonlinear system subjected to LIM with small values of τ_0, as shown below.

Consider a sample matrix

$$\mathbf{G}'(\tau_0) = \langle \boldsymbol{x}(t+\tau_0)\boldsymbol{x}^{\mathrm{T}}(t)\rangle\langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle^{-1} \tag{27}$$

that is *not* well approximated by a matrix exponential. We do assume that it is smooth enough to be Taylor expanded in an absolutely convergent series for small values of τ_0:

$$\begin{aligned}\mathbf{G}'(\tau_0) &= \mathbf{1} + \mathbf{A}'\tau_0 + \mathbf{A}''\tau_0^2 + \ldots \\ &= \mathbf{1} + \mathbf{A}'\tau_0 + (\mathbf{A}'\tau_0)^2/2 + (\mathbf{A}''-\mathbf{A}'^2/2)\tau_0^2 + \ldots \\ &= \exp(\mathbf{A}'\tau_0) + O(\tau_0^2) \\ &\approx \exp([\mathbf{A}'+\mathbf{a}\tau_0]\tau_0). \end{aligned} \tag{28}$$

In Eq.(28), the symbol $O(\tau_0^2)$ indicates all terms of at least second order in τ_0, and the matrix **a** is the difference between the second term in the Taylor expansion of $\mathbf{G}'(\tau_0)$ and that of a matrix exponential. Thus, we identify the slope (-140.3/model unit^2) of the curve at small values of τ_0 in Fig. 27.3 as trace(**a**).

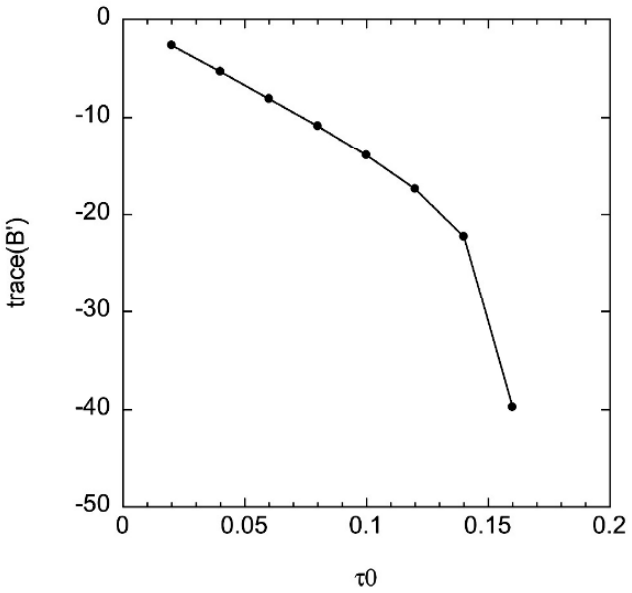


Fig. 27.3. Trace(**B'**) *vs.* τ_0 for LIM applied to L63.

It was shown in Penland and Ghil (1993) that a similar dependence on τ_0 obtains when the system is linear but the observations do not sufficiently span the dynamical space. This is an indication of the well-known mathematical equivalence between nonlinear models and linear models of larger dimensionality. If the system in fact obeys dynamics with negligible nonlinearity and the observations do not span the dynamical space, one must obviously generate the unobserved directions with a nonlinear model in order to fit the data. Conversely, one can usually fit a nonlinear model with a linear model of sufficiently high dimensionality by redefining temporal derivatives. In that case, the number of dimensions needed to fit the data can provide valuable information on the character of the nonlinearity in physical space. As humans, we are anxious to ask, "But what is it *really*?" Heartless Nature responds that it depends on how we choose (or are able) to look at it.

Returning to the Lorenz system, it may be hoped that linear predictions may be useful even though we know from Fig. 27.3 that the system is nonlinear in the observed space. Fig. 27.4 obliterates this hope. The observed error variance (solid lines), normalized to the total variance of the 3-dimensional time series, show deviations from the expected error variance (trace($\sigma(\tau)$, see Eq. 20)) for all values of τ_0. Further, the curves are strongly dependent upon the value of τ_0, with $\tau_0 = 0.02$ (open circles) estimating much *smaller* prediction error variance than $\tau_0 = 0.16$ (heavy solid line) but delivering much *larger* errors. Although Fig. 27.4 reaches the same conclusion as PS95 (their Fig. 19), the figures themselves look somewhat different, even though both studies considered the same Lorenz system, using the

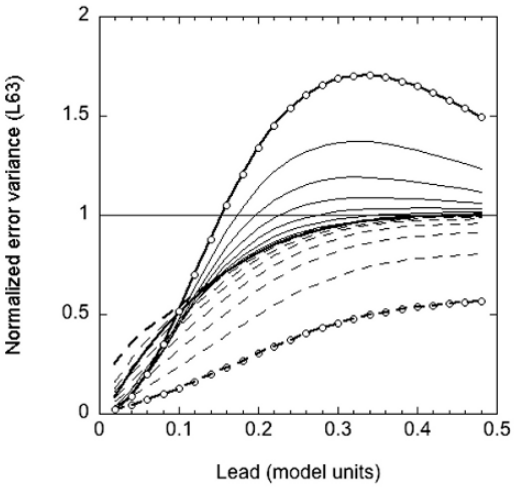


Fig. 27.4. Variance of linear prediction error, normalized to the total variance of the Lorenz system. Dashed lines: Theoretically expected error variance predicted by LIM. Solid lines: True error variance. Plain heavy solid line: $\tau_0 = 0.16$ model units. Heavy solid line with open circles: $\tau_0 = 0.02$. Light solid lines correspond to the other values of τ_0.

same values of τ_0, and for the same lead times. The resolution of this apparent paradox is that PS95, in an attempt to mirror the treatment of their sea surface temperature data set as closely as possible, used a verification time series much shorter than the one we used here. It is therefore possible that their verification time series sampled only one lobe of the attractor. If the linear approximation hypothesis were correct, of course, there would no difference. It is also noted that the PS95 study carefully distinguished between training period and verification period. Here, we have used the entire length of the time series for both training and verification since we are more interested in whether the curves vary with τ_0 rather than the actual size of the error variance.

Note that the observed error variance can be larger than the saturation value of unity, particularly for small values of τ_0. The normalized error variance of the persistence forecasts (not shown) is also larger than the saturation value; its peak value of two occurs at a lead of about 0.4 units.

Finally, one may ask if a coarse-grained version of L63 may be treated as a linear stochastic process. Starting with the sampling interval, we smoothed the L63 time series on progressively longer time scales and were unable to find any coarse-graining that gave a useful linear description of the system. The oscillations of the L63 are rapid enough that Nyquist problems occurred before the linear regime could be reached. That is, oscillations were too rapid to be resolved at the smallest τ_0 allowed by the coarse-graining, resulting in the imaginary part of an eigenvalue of $\mathbf{B}'\tau_0$ equal to $n\pi$, where n is a nonzero integer. We conclude that any possible three-dimensional linear regime of L63 would be at a scaling coarse enough that oscillations around the center of a lobe are unresolved.

3.2 A linear process coupled to the chaotic Lorenz system

A situation is now considered where the Lorenz system is considered as external forcing for a linear system. That is, we wish to investigate whether a modest coarse-graining allows the L63 to be treated as a stochastic component. Redefining x, y, and z,

$$\begin{aligned} \frac{dx}{dt} &= \alpha x + l_x \\ \frac{dy}{dt} &= \beta y + l_y \\ \frac{dz}{dt} &= \gamma z + l_z \end{aligned} \quad . \tag{29}$$

In Eq.(29), $\boldsymbol{l}$ is defined as in Eq.(26); that is, we use the centered Lorenz system to force the components $\boldsymbol{u} = (x, y, z)$. The values of (α, β, γ) are (-0.02, -0.033, -0.05). Eq. (29) is integrated using a time step of 0.02 model units, i.e., the sampling of the Lorenz system shown in Fig. 27.1. The result is coarse-grained on intervals of two model units, so that each recorded component of $\boldsymbol{u}$ is an average over 100 samples.

Because the model has been coarse-grained over two model units, we shall call this case Model 2. Note that we rewrite Eq. (29) as

$$\frac{d\boldsymbol{u}}{dt} = \mathbf{B}\boldsymbol{u} + \boldsymbol{l}, \tag{30}$$

where, we repeat, $\boldsymbol{u}$ and $\boldsymbol{l}$ have been coarse grained over 100 time steps, and the matrix $\mathbf{B}$ is diagonal with trace($\mathbf{B}$) = -0.103.

LIM was applied to the coarse grained time series and samples $\mathbf{B}'(\tau_0)$ estimated for τ_0 = 2, 4, 6 , 8, 10, 12, 14 and 16 model units. For values of τ_0 larger than 16, Nyquist problems occurred. The trace($\mathbf{B}'$) is shown in Fig. 27.5 (solid circles), along with the trace (-0.103) expected if the L63 system truly acted as stationary white noise forcing on those timescales. Clearly, $\mathbf{B}'$ does not perfectly pass the tau test, particularly for values of τ_0 smaller than 6. Nevertheless, even for $\tau_0 \leq 6$, the slope (0.013/model unit^2) is four orders of magnitude smaller than that of the linearized regime of L63. This is less impressive than it sounds, but not much. One should actually compare that slope with the trace of the average trace($\mathbf{B}'$). For L63, that ratio is about -20; for the system Eq. (29), coarse grained at 2 model units, the ratio is closer to 0.1.

We also coarse-grained $\boldsymbol{u}$ over 200 time steps, equivalent to 4 model units. Hereafter, this case will be known as Model 4. We can see from Fig. 27.1 that 4 model units do not allow sufficient time for many passages between lobes. Nevertheless, the doubling of the averaging time does seem to improve the linear approximation in that the slope of trace($\mathbf{B}'$) with time is greatly reduced, and the values are closer to the expected value of -0.103.

From Table 27.1, we see that the reproduction of $\mathbf{B}$ by the estimates $\mathbf{B}'$ is imperfect, but does at least suggest that $\mathbf{B}$ is dominated by its diagonal elements. The high correlation of l_x and l_y (0.85) causes LIM to confuse the linear behavior of x and y

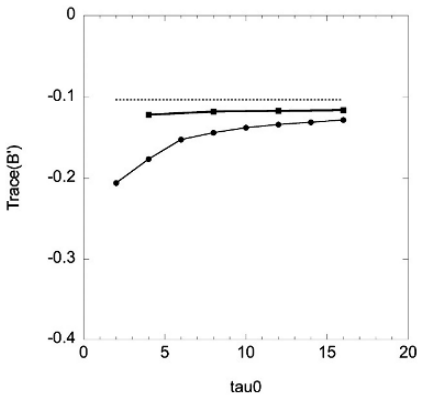


Fig. 27.5. Trace($\mathbf{B}'$) for Eq. (28). Light solid line: Model 2. Heavy solid line: Model 4. Dotted line: Trace($\mathbf{B}'$)= – 0.103/model unit.

somewhat. Nevertheless, the amount of damping introduced by **B** is approximately the right size, with the coarse graining, performed after the model integration, introducing a bit of extra damping to z.

Table 27.1

Empirical estimates of the linear operator in Eq. (28)

Model 2:

$$\mathbf{B'}_{ave} = \begin{pmatrix} -0.023 \pm 0.016 & 0.006 \pm 0.026 & 0.008 \pm 0.010 \\ -0.001 \pm 0.015 & -0.031 \pm 0.023 & 0.007 \pm 0.011 \\ 0.002 \pm 0.0004 & -0.002 \pm 0.0006 & -0.097 \pm 0.033 \end{pmatrix}$$

Model 2($\tau_0 \geq 6$):

$$\mathbf{B'}_{ave} = \begin{pmatrix} -0.014 \pm 0.002 & -0.008 \pm 0.004 & 0.001 \pm 0.004 \\ 0.008 \pm 0.002 & -0.044 \pm 0.003 & 0.0002 \pm 0.004 \\ 0.001 \pm 0.0002 & -0.002 \pm 0.0002 & -0.078 \pm 0.007 \end{pmatrix}$$

Model 4:

$$\mathbf{B'}_{ave} = \begin{pmatrix} -0.026 \pm 0.014 & 0.012 \pm 0.022 & 0.014 \pm 0.012 \\ -0.004 \pm 0.013 & -0.024 \pm 0.021 & 0.013 \pm 0.012 \\ 0.001 \pm 0.0002 & -0.002 \pm 0.0003 & -0.069 \pm 0.010 \end{pmatrix}$$

Integrated Model:

$$\mathbf{B} = \begin{pmatrix} -0.020 & 0.000 & 0.000 \\ 0.000 & -0.033 & 0.000 \\ 0.000 & 0.000 & -0.050 \end{pmatrix}$$

The numerical values for **B'** and **Q'** as estimated by the following: 1) Model 2, averaged over values of τ_0 = 2, 4, 6, 8, 10, 12, 14 and16 model units; 2) again Model 2, but averaging over only τ_0 = 8, 10, 12, 14 and16 model units; i.e., the linear regime as indicated by the flattening of trace(**B'**) shown in Fig. 27.5; and 3) Model 4, averaged over values of τ_0 = 4, 8, 12 and 16 model units. Table 1 gives the averaged values of B'_{ij}; Table 27.2 gives the averaged values of Q'_{ij} along with the contemporaneous covariance matrix and correlation matrix of L63. The uncertainty estimates are not those associated with the length of the time series, but are rather indicative of the variation with τ_0.

Table 27.2

Empirical estimates of the forcing covariance matrix

Model 2:

$$\mathbf{Q}'_{ave} = \begin{pmatrix} 53.14 \pm 7.39 & 53.45 \pm 7.35 & 0.12 \pm 0.38 \\ 53.45 \pm 7.35 & 53.86 \pm 7.27 & 0.13 \pm 0.04 \\ 0.12 \pm 0.38 & 0.13 \pm 0.04 & 1.21 \pm 0.42 \end{pmatrix}$$

Model 2 ($\tau_0 \geq 6$):

$$\mathbf{Q}'_{ave} = \begin{pmatrix} 57.07 \pm 0.86 & 57.37 \pm 0.90 & 0.12 \pm 0.05 \\ 57.37 \pm 0.90 & 57.74 \pm 0.92 & 0.14 \pm 0.05 \\ 0.12 \pm 0.05 & 0.14 \pm 0.05 & 0.97 \pm 0.09 \end{pmatrix}$$

Model 4:

$$\mathbf{Q}'_{ave} = \begin{pmatrix} 47.88 \pm 6.88 & 48.22 \pm 6.95 & 0.02 \pm 0.03 \\ 48.22 \pm 6.95 & 48.56 \pm 6.99 & 0.03 \pm 0.03 \\ 0.02 \pm 0.03 & 0.03 \pm 0.03 & 0.73 \pm 0.10 \end{pmatrix}$$

Covariance matrix of L63:

$$\langle \boldsymbol{l}\boldsymbol{l}^{\mathrm{T}} \rangle = \begin{pmatrix} 95.49 & 95.32 & -0.26 \\ 95.32 & 131.07 & -0.33 \\ -0.26 & -0.33 & 107.14 \end{pmatrix}$$

Correlation matrix of L63:

$$\mathbf{C}_{L63} = \begin{pmatrix} 1.00 & 0.852 & -0.003 \\ 0.852 & 1.00 & -0.003 \\ -0.003 & -0.003 & 1.00 \end{pmatrix}$$

The effect of the coarse graining is particularly noticeable in Table 27.2. Although the estimates $\mathbf{Q}'$ are not directly related to $\langle \boldsymbol{l}\boldsymbol{l}^{\mathrm{T}} \rangle$ (they have different units, for one thing, since the white noise approximation to L63 would be a complicated time integral over its lagged covariance structure), a surprising amount of information about L63 is obtained in this way. The high correlation between l_x and l_y is reproduced, and each is shown to be only weakly correlated with l_z. That the value of Q'_{zz} is severely reduced by the coarse graining is clear from Fig. 27.1; the variation in the

coarse-grained l_x and l_y comes primarily from jumping between the lobes of the attractor while most of the l_z variation is simply averaged out.

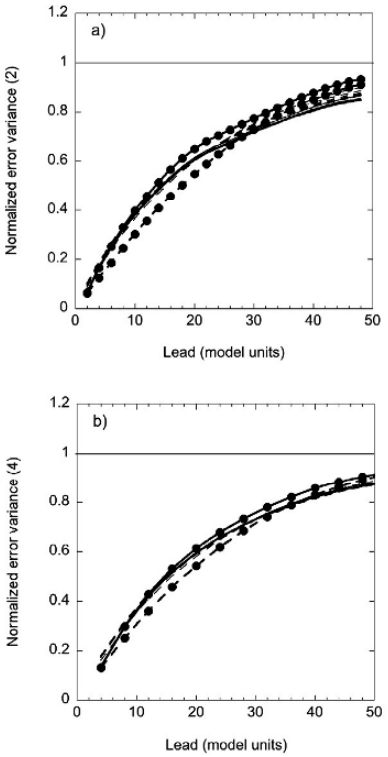


Fig. 27.6. Error variance curves, normalized to the field variance. Dashed and solid lines as in Fig. 27.4. Filled circles: smallest value of τ_0. Heavy line with no symbols: largest value of τ_0. *a*) Model 2. *b*) Model 4.

In terms of linear predictability, we see from Fig. 27.6 that the linear models of both coarse-grained systems pass the tau test reasonably well, though imperfectly. One notes in each case that the smallest value of τ_0 (filled circles) promises the best, and delivers the worst, prediction skill. Nevertheless, the spread of prediction error with τ_0 compared with that in Fig. 27.4 is very small, providing quantitative evidence for the validity of linearly approximating these coarse-grained systems.

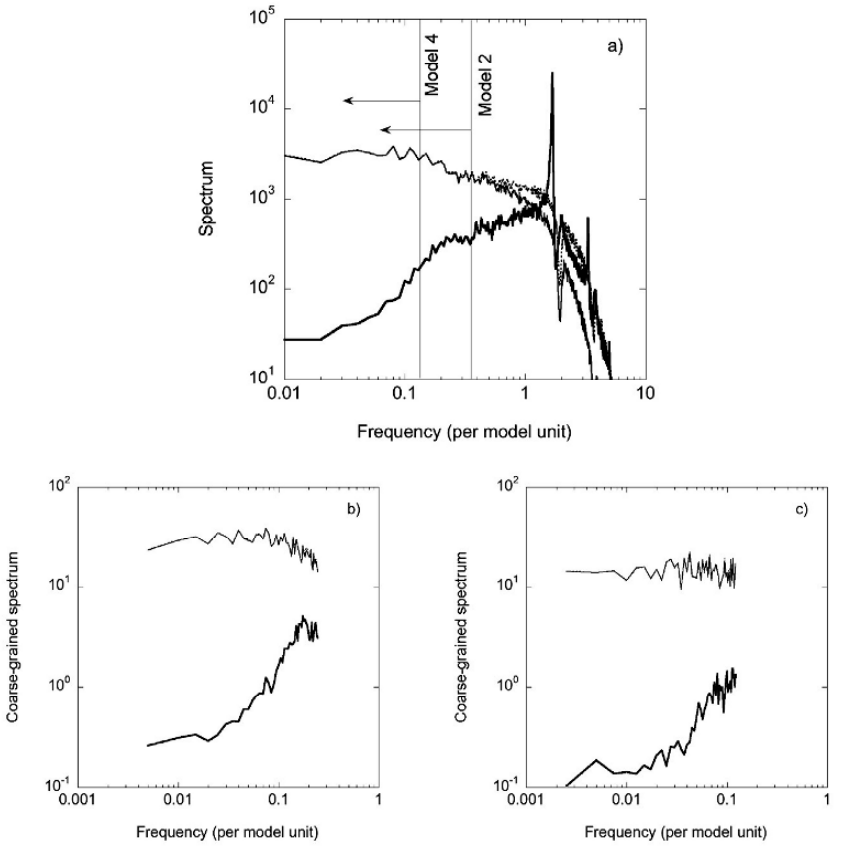


Fig. 27.7. Fast Fourier Transform spectra for l_x (light solid line), l_y (dotted line) and l_z (heavy solid line) components of the chaotic Lorenz system. *a*) Original system. *b*) Model 2. *c*) Model 4. At low frequencies, the spectra for l_x and l_y cannot be distinguished.

As a final diagnosis of the validity of treating the coarse-grained Lorenz model as white noise in Eq. (30), we present spectra of l_x, l_y and l_z in Fig. 27.7 corresponding to the original Lorenz system sampled at 0.02 model units (Fig. 27.7a), Model 2 (Fig. 27.7b) and Model 4 (Fig. 27.7c). In Fig. 27.7a, we have denoted the frequencies resolved by Models 2 and 4 for comparison with spectra in Figs. 27.7b and 27.7c. It is clear that the coarse-graining does more than change the resolution; it also affects both the magnitude and the shape of the spectra. While it is no surprise that the sharp peaks in the spectrum of l_z are eliminated, one might remark that the amplitude of the broadband part of that spectrum is also reduced by orders of magnitude. The same is true of l_x and l_y. This property underscores the importance of $1/\varepsilon$ in the scaled Eq. (4);

if the rapidly varying contribution to the underlying dynamical system is not large enough, coarse-graining will simply average out its effects. Further, for the white noise approximation to be valid in the coarse-grained system, the high-frequency band of the original spectrum must have a broad enough nature that at least some of that variance can be aliased into the low frequencies by the coarse graining, thus flattening out the spectrum. This is the case for l_x and l_y, whose spectra in the coarse-grained system are virtually identical. However, the component l_z of the original system has much of its variance in nearly sinusoidal behavior that the coarse-graining simply averages out.

4 Experiments Using Tropical Sea Surface Temperatures

In our examples using linear stochastic approximations so far, we have greatly profited from the blessings of knowing in advance the architecture of the underlying dynamical system. When we applied LIM to coarse-grained models, we had a "truth" with which to compare our answers. When we considered the spectra of the coarse-grained forcing, we had a time series of that forcing to analyze. This is not the case when the dynamics of natural phenomena are diagnosed from observed time series, which is why so many researchers like to start with a model "having many features of the observed system" and assume that it describes reality.

One purpose of this manuscript is to present one method for quantifying the "many features" statement. The accuracy with which observations can be described by a linear system, even when the underlying dynamics are chaotic, is a physical characteristic of the phenomenon. Comparison between LIM applied to data and LIM applied to a model can therefore aid in judging the fidelity of the model to reality. We believe this procedure could be a valuable tool in evaluating the enormous number of models claiming to represent the dynamics of tropical sea surface temperatures related to "El Niño," described below.

Originally known as a phenomenon occurring off the western coast of tropical South America, the term "El Niño" now generally indicates an episodic warming of sea surface temperatures (SSTs) over most of the tropical Pacific. With the availability of longer and better data sets, it is gradually becoming obvious that the "El Niño signal" is a global phenomenon, with clear signatures in data such as south tropical Atlantic SSTs (Mo and Häkkinen 2001; Penland and Matrosova 2006). Multiscale interactions are crucial to the evolution of El Niño in every theory proposed to explain it.

Unfortunately for diagnostics, salient features of El Niño can be found in models based on every possible genre of attractor, whether they be strange attractors (e.g., Jin et al. 1994; Tziperman et al. 1994), limit cycles or quasi-limit cycles (e.g., Battisti and Hirst 1989; Jin and Neelin 1993*a,b*; Neelin and Jin 1993; Blanke et al. 1997; Sun et al. 2004; Sun 2007) or fixed points (e.g., Penland and Magorian 1993; PS95; Moore and Kleeman 1996; Kleeman and Moore 1997; Thompson and Battisti 2000, 2001). Simple conceptual models of El Niño are therefore unable to give dynamical clarification of what the El Niño attractor looks like.

Linear theories of El Niño invoke the nonnormality of the deterministic linear operator (e.g., Farrell 1988; PS95) to explain the growth of SST anomalies during the development of an El Niño or, in its opposite phase, La Niña event.

As we have seen, there may be a timescale on which the paradigm of a chaotic attractor and that of a fixed point are mutually consistent. This statement should be taken in the spirit of the statement that the classical dynamics, the ideal gas law for example, is consistent with quantum mechanics. There is much evidence that on seasonal timescales El Niño dynamics reside in the linear regime and that the chaotic nonlinear dynamical component may be treated as stochastic forcing, with variance depending on the annual cycle. For modeling evidence, we refer the reader to works by R. Kleeman, and A. Moore, and the later (post-2000) modeling studies of D. Battisti. Some of the observational evidence is summarized here.

As stated in Section 2, linear theories of El Niño invoke the nonnormality of the deterministic linear operator to explain the growth of SST anomalies during the development of an El Niño or, in its opposite phase, La Niña event. One dominant optimal initial structure for growth is found when LIM is applied to tropical SSTs, and this structure evolves into a mature El Niño event about 6-9 months after its appearance (Fig. 27.8). The optimal structure does not significantly change with values of τ_0 (Penland 1996). PS95 and Penland (1996) confined their studies to the tropical Indo-Pacific, but inclusion of the entire tropical belt (Penland and Matrosova 1998, 2006) only makes their conclusions more robust. In fact, using the constituent normal modes to isolate the El Niño signal (Penland and Matrosova 2006) underscored the global nature of El Niño by revealing that its signal in the north tropical Atlantic ocean is very similar to the El Niño signal in the central Indian ocean (contemporaneous correlation +0.8), and that a strong precursor to SST

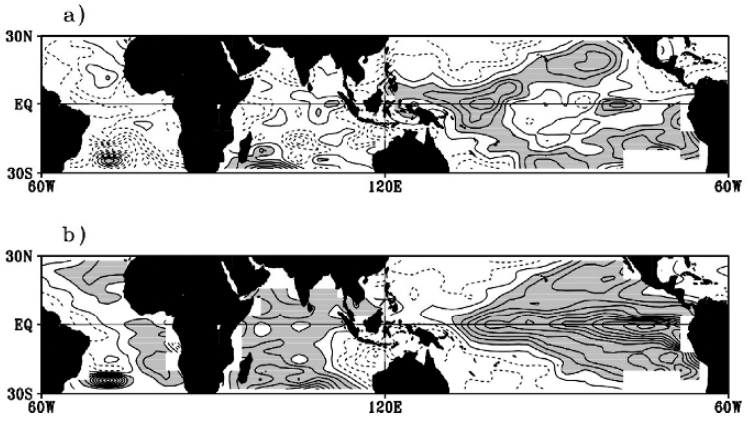


Fig. 27.8. Optimal structure of tropical SSTA and b) its evolution in eight months. The contour interval in the domain 60W-120E is 0.015. The contour interval in the domain 120E-60W is 0.03. Areas enclosed by the first positive contour are shaded. Dotted lines represent negative contours.

anomalies in the traditional Niño 3.4 region (6°N-6°S, 170°W-120°W) is found in the tropical south Atlantic ocean (lagged correlation, S. Atlantic leading, -0.6). This precursor occurs because the projection onto the optimal structure is so strong in the south tropical Atlantic, so using SSTs in that region to predict El Niño is no better than using the projection of the optimal structure as a whole. These global teleconnections indicate that any credible theory of El Niño cannot be based on a single time series such as the Niño 3.4 SST anomaly or the Southern Oscillation Index alone; such a theory must also be able to predict the global El Niño signal.

3.1 The data set

We consider COADS monthly SST data (Woodruff et al. 1993) in the entire tropical strip between 30°N and 30°S. The data were consolidated onto a $4^o \times 10^o$ grid and subjected to a three-month running mean. The 1950-2000 climatology was then removed from the SST data and the anomalies were projected onto 20 Empirical Orthogonal Functions (EOFs: Hotelling 1933) containing about two-thirds of the variance. Obviously, massaging the data as just described is tantamount to filtering the data. However, for the purposes of isolating the El Niño signal in this study, we shall refer to this data set as the "unfiltered data."

LIM was applied to the unfiltered data using τ_0 = 4 months. The usual optimal structure was found to project onto three complex normal mode pairs associated with periods of about 2 years, 5 years, and 20 years. The dominance of these three modal pairs in the nonnormally evolving El Niño signal was found to be robust for those values of τ_0 uncompromised by Nyquist problems. The global El Niño signal was extracted from the multivariate data set by projecting the data onto the adjoint patterns of the three modal pairs. As shown by Penland and Sardeshmukh (1995*a*), the uncertainty in these timescales is large. However, Penland and Matrosova (2006) showed that the data filtered in this way accounted for the relevant Fourier spectral frequencies, as well as most of the low-frequency variability of the data set. The red noise background remained in the residual data.

To show that isolating the El Niño signal this way is consistent with the traditional Niño 3.4 SST anomaly index, we show both the filtered and unfiltered versions of that index in Fig. 27.9. The two indices are highly correlated (correlation

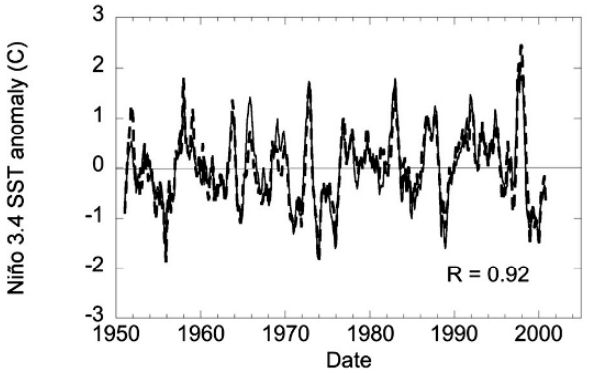


Fig. 27.9. Time series of Niño 3.4 SST anomaly. Solid line: unfiltered data. Dashed line: El Niño pass data.

+0.92), and mostly indistinguishable from each other. If one accepts the National Oceanic and Atmospheric Administration's designation of the unfiltered Niño 3.4 SST anomaly index as the official El Niño indicator, this check verifies that the filter, which is a pattern-based method, captures most of the El Niño variability. Although the modal and adjoint patterns were derived using an assumption of linear dynamics, there is no reason to believe that a map-by-map projection onto these time-independent patterns would eliminate any important nonlinear effects in the temporal evolution of El Niño, particularly since there are no large differences between the filtered and unfiltered versions of this index. We now proceed with the six-dimensional El Niño-pass filtered data set.

3.2 Results for SSTs

Our first order of business in treating the filtered data is to project them onto their EOFs, of which there are six. A detailed description of these EOFs and their time-dependent coefficients is found in Penland and Matrosova (2006). LIM was applied to this system and estimates $\mathbf{B}'$ of the linear operator were obtained using values of τ_0 = 3, 4, 5, 6, 7, 8, 9 and 10 months. Nyquist problems appeared for larger values of τ_0. Interestingly, Nyquist problems did not appear for τ_0 near 6 months, in agreement with Penland's (1996) suggestion that the well-known dependence of El Niño on the annual cycle is mainly due to the yearly varying variance of the fast processes acting as stochastic forcing. As long as statistics are accumulated over an integral number of years, a periodically varying $\mathbf{S}$ in Eq. (14) can easily be handled by LIM; the only modification occurs in interpreting the matices $\langle \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\rangle$ and $\mathbf{Q}$ in Eq. (17) as annual averages.

A plot of trace($\mathbf{B}'$) *vs.* τ_0 is shown in Fig. 27.10. The dependence on τ_0 is linear for values of τ_0 less than about 6 months, after which the curve flattens out. The slope before the break is about 0.04 (months)$^{-2}$. Interestingly, this result implies a role for nonlinearities somewhat slower than that found by PS95 (their Fig. 12).

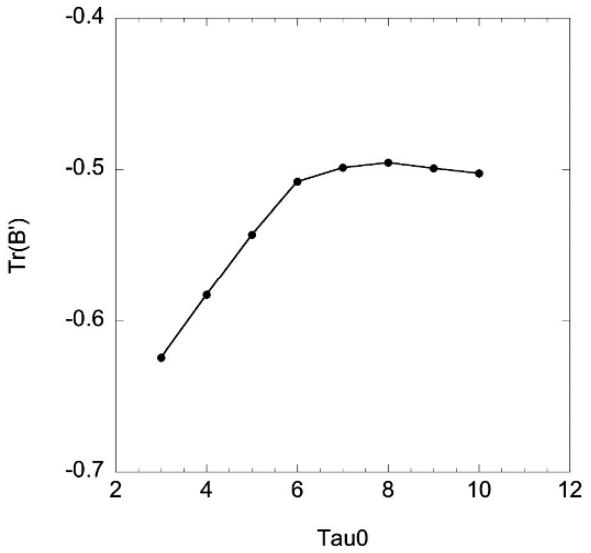


Fig. 27.10. Trace(**B'**) *vs.* τ_0 for tropical SST anomalies.

Of course, their 15-dimensional system contained dynamical process other than El Niño, and their graph was contaminated with Nyquist effects. However, the two studies agree in that the linear approximation is fairly good, particularly at timescales larger than 6 months.

A similar story is told by the error variance of the linear predictions (Fig. 27.11). Unlike the linear system forced by L63, the worst predictions are not at the smallest value of τ_0, but rather at the largest (10 months). This may occur because of nonlinearities, or it may be due to proximity to the half period of the quasibienniel oscillation. In any case, it is encouraging to see that the error variance for $\tau_0 = 4$ months is not significantly larger than that for $\tau_0 = 6$ months, in spite of Fig. 27.10, since forecasts currently provided to the general public from the website http://www.cdc.noaa.gov were developed with $\tau_0 = 4$ months.

Nevertheless, the variation with τ_0 in Fig. 27.11 is not as tight as that found by PS95. It appears that application of the nonnormal filter, by eliminating much of the red noise background associated with other linear processes in tropical SST, may actually magnify the presence of nonlinearity in the El Niño signal. This does not mean that nonlinear dynamics dominate its evolution on the seasonal scale. On the contrary, comparing the results of LIM applied to SST with those of the L63 results verifies the usefulness of the linear approximation for tropical SST anomaly dynamics.

The estimates **B'** for the 6-dimensional system spanned by the leading EOFs of the El Niño-pass SST data are provided in Table 27.3. Estimates **Q'** are provided in Table 27.4. It is encouraging that **Q'** is positive definite, a situation not at all obvious with the linear system driven by the Lorenz model.

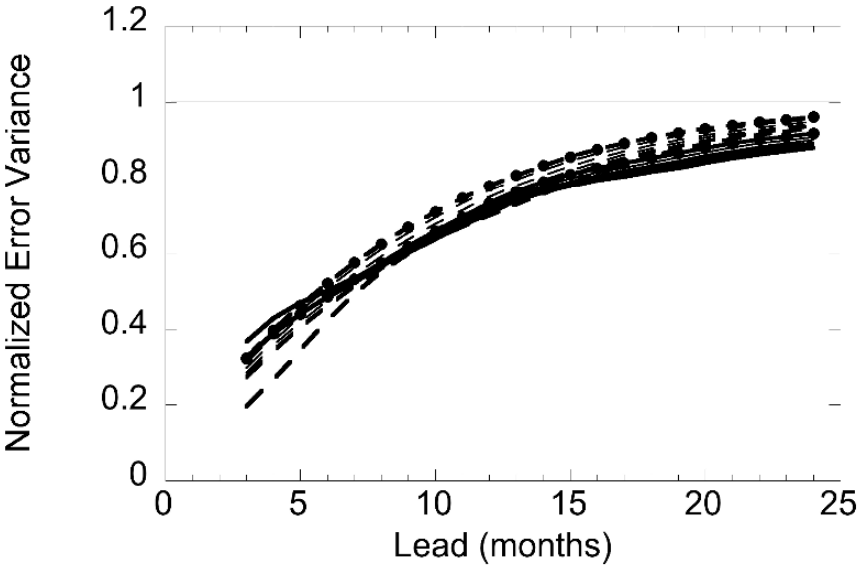


Fig. 27.11. Error variance normalized to SST anomaly field variance. Conventions as in Fig. 27.6.

Table 27.3

Empirical estimates of the linear operator from tropical SST

All values of τ_0:

B' =

$$\begin{pmatrix} -0.03 \pm 0.001 & 0.05 \pm 0.002 & -0.02 \pm 0.001 & 0.02 \pm 0.003 & 0.02 \pm 0.001 & -0.03 \pm 0.001 \\ -0.11 \pm 0.001 & -0.05 \pm 0.01 & 0.03 \pm 0.004 & -0.03 \pm 0.002 & 0.01 \pm 0.005 & -0.01 \pm 0.002 \\ 0.11 \pm 0.002 & 0.07 \pm 0.004 & -0.09 \pm 0.01 & -0.13 \pm 0.01 & -0.08 \pm 0.003 & 0.02 \pm 0.005 \\ -0.21 \pm 0.005 & 0.04 \pm 0.01 & 0.17 \pm 0.004 & -0.15 \pm 0.02 & 0.13 \pm 0.009 & -0.06 \pm 0.002 \\ -0.13 \pm 0.004 & -0.04 \pm 0.007 & 0.05 \pm 0.001 & -0.20 \pm 0.01 & -0.15 \pm 0.005 & -0.09 \pm 0.005 \\ 0.18 \pm 0.006 & 0.045 \pm 0.002 & -0.05 \pm 0.001 & 0.06 \pm 0.01 & 0.08 \pm 0.002 & -0.08 \pm 0.01 \end{pmatrix}$$

Values of $\tau_0 \geq 6$ months:

B' =

$$\begin{pmatrix} -0.03 \pm 0.001 & 0.05 \pm 0.001 & -0.02 \pm 0.01 & 0.03 \pm 0.01 & 0.01 \pm 0.002 & -0.03 \pm 0.003 \\ -0.11 \pm 0.002 & -0.04 \pm 0.004 & 0.03 \pm 0.002 & -0.03 \pm 0.004 & 0.01 \pm 0.002 & -0.01 \pm 0.003 \\ 0.078 \pm 0.05 & -0.07 \pm 0.01 & -0.06 \pm 0.03 & -0.16 \pm 0.03 & -0.08 \pm 0.01 & 0.01 \pm 0.01 \\ -0.27 \pm 0.03 & 0.04 \pm 0.02 & 0.17 \pm 0.01 & -0.16 \pm 0.05 & 0.10 \pm 0.03 & -0.06 \pm 0.001 \\ -0.18 \pm 0.09 & -0.03 \pm 0.003 & 0.07 \pm 0.04 & -0.26 \pm 0.08 & -0.14 \pm 0.01 & -0.11 \pm 0.02 \\ 0.18 \pm 0.006 & 0.19 \pm 0.01 & 0.04 \pm 0.01 & -0.05 \pm 0.01 & 0.06 \pm 0.01 & -0.07 \pm 0.01 \end{pmatrix}$$

Table 27.4

Empirical estimates of the stochastic forcing from tropical SST

All values of τ_0:

Q' =

$$\begin{pmatrix} 1.52 \pm 0.50 & -0.28 \pm 0.31 & -0.19 \pm 0.21 & 0.15 \pm 0.81 & 0.09 \pm 0.27 & 0.14 \pm 0.17 \\ -0.28 \pm 0.31 & 1.07 \pm 0.15 & 0.22 \pm 0.08 & -0.04 \pm 0.27 & -0.01 \pm 0.14 & -0.13 \pm 0.21 \\ -0.19 \pm 0.21 & 0.22 \pm 0.08 & 1.13 \pm 0.42 & 0.11 \pm 0.26 & 0.34 \pm 0.08 & 0.02 \pm 0.03 \\ 0.15 \pm 0.81 & -0.04 \pm 0.27 & 0.11 \pm 0.26 & 1.52 \pm 0.29 & 0.19 \pm 0.21 & 0.10 \pm 0.10 \\ 0.09 \pm 0.27 & -0.01 \pm 0.14 & 0.34 \pm 0.08 & 0.19 \pm 0.21 & 1.43 \pm 0.07 & 0.08 \pm 0.22 \\ 0.14 \pm 0.17 & -0.13 \pm 0.21 & 0.02 \pm 0.03 & 0.10 \pm 0.10 & 0.08 \pm 0.22 & 0.52 \pm 0.20 \end{pmatrix}$$

Values of $\tau_0 \geq 6$ months:

Q' =

$$\begin{pmatrix} 1.55 \pm 0.50 & -0.16 \pm 0.04 & 0.05 \pm 0.54 & 0.44 \pm 0.08 & 0.47 \pm 0.45 & 0.19 \pm 0.13 \\ -0.16 \pm 0.04 & 0.90 \pm 0.08 & 0.19 \pm 0.13 & 0.11 \pm 0.16 & 0.04 \pm 0.04 & -0.03 \pm 0.01 \\ 0.05 \pm 0.55 & 0.19 \pm 0.06 & 0.83 \pm 0.43 & 0.24 \pm 0.23 & 0.24 \pm 0.23 & 0.11 \pm 0.12 \\ 0.43 \pm 0.08 & 0.11 \pm 0.16 & 0.24 \pm 0.23 & 1.79 \pm 0.61 & 0.68 \pm 0.52 & 0.10 \pm 0.04 \\ 0.47 \pm 0.45 & 0.04 \pm 0.04 & 0.24 \pm 0.23 & 0.68 \pm 0.52 & 1.40 \pm 0.08 & 0.28 \pm 0.13 \\ 0.19 \pm 0.13 & -0.03 \pm 0.01 & 0.11 \pm 0.12 & 0.10 \pm 0.04 & 0.28 \pm 0.13 & 0.53 \pm 0.06 \end{pmatrix}$$

4 Conclusions

It is true, generally speaking, that the universe is governed by nonlinear dynamics. It is also true, generally speaking, that the universe is governed by quantum mechanics. In fact, given the ubiquity of electromagnetic radiation, vacuum fluctuations, and interstellar media comprising atomic and subatomic particles, not to mention the relatively small amount of space taken up by bright matter, there is a case to be made that the subset of dynamical systems for which the classical approximation obtains is a very small set indeed. Nevertheless, for us who are confined to a terrestrial environment, that small set is extremely important. Further, there is a grey area, particularly in molecular dynamics, where hybrid models, basically using classical dynamics but incorporating quantum mechanical effects with relevant approximations, have proved themselves as valuable research tools. The relationship between classical and quantum mechanics as an allegory for that between linear and nonlinear dynamics is not exact, but it is useful to remind us that there is a very large grey area where the dynamical system may be treated as basically linear, although nonlinearities may not be ignored.

The importance of weather and climate predictions to the well-being of terrestrial inhabitants demands that meteorologists and climatologists provide timely forecasts

of the highest possible quality. Linearly based forecasts may be timely and easy to make, but no credible scientist would advocate using them for these reasons alone. However, they *are* easy to make and verify, and are sometimes even the best forecasts available. For example, LIM forecasts of river flows in Colombia, a system this author would never have the courage to consider linear, have shown themselves to be a skillful component of estimating electricity needs in that nation (Germán Poveda 2002, personal communication). It therefore behooves us to examine quantitatively the extent to which the linear approximation is valid. Recall that this approximation does not ignore nonlinearities; rather, it is valid when they vary rapidly enough to be treated as a stochastic component, possibly modulated by a linear term.

We have reviewed a procedure for finding the optimal linear model of a multivariate process, along with quantitative criteria for assessing its validity (the tau test). We have also briefly reviewed some of the limit theorems (e.g., Khasminsii 1966; Papanicolaou and Kohler 1974) that are relevant to the part of the approximation that treats the rapidly varying system as a stochastic term. There are later articles extending the limit theorems, notably by Kifer (2001). Further, the decomposition of a chaotic attractor into Markov partitions, and the resulting probabilistic description of those systems, is discussed by Dorfman (1999) and references therein. Our approach here is not theoretical rigor as much as it is to give a flavor of when the linear approximation may be useful. This is done in the context of three systems: the low-order chaotic Lorenz model (L63), a coarse-grained linear system for which L63 was treated as external forcing, and observed seasonal sea surface temperature anomalies.

It is not surprising that a linear approximation of L63 yields poor predictions of that system, or that tests for validity of the linear approximation fail spectacularly. In fact, this failure may be compared with the failure of LIM applied to another low dimensional system by Tsonis et al. (2006). The usefulness of these exercises lies in their providing benchmarks for how highly nonlinear, low-order chaotic systems perform on such tests.

Our second experimental set-up consisted of a linear system with decay times long compared with the timescales of L63 (Fig. 27.1), which maintained the linear system by acting as an external forcing. Although the linear decay times were long, the coarse-graining and sampling interval of 2 model units (Model 2) was purposely chosen to resolve some of L63's strong nonlinearity in order to strain the linear approximation. We compared Model 2 with a model coarse-grained and sampled at 4 model units (Model 4). Again, our expectations were confirmed when Model 4 passed the tau test better than Model 2 did. On the other hand, Model 2 gave us more information about the system: it told us the validity of the linear system was probably best for coarse-grainings greater than 6 model units, it told us the relative size, compared to linear dynamics, of the nonlinearities for coarse-grainings between 2 and 6 model units, and it gave just as good insight as Model 4 into the relative variance and correlation of the three L63 components.

It is encouraging that many characteristics of L63 could be inferred from a linear fluctuation-dissipation relation (for the purists: "of the second kind", Cugliandolo, Dean and Kurchan 1997) derived from Models 2 and 4. That it is able to diagnose

high correlations, such as that between l_x and l_y, suggests that a similar analysis of numerical model errors might reveal unsuspected relations between different sources of error.

Analysis of tropical sea surface temperatures basically confirmed the results of PS95, who found a linear approximation of their dynamics to be good and useful, though imperfect. A dynamically based filter (Penland and Matrosova 2006) used to isolate the El Niño signal seemed to unmask a small nonlinear component to this signal, even though the filter itself was based on an assumption of linear dynamics. Of course, Fourier filtering is useful even when it falsely assumes that the system is perfectly periodic with a period equal to the length of the time series; similarly, the assumption of linear dynamics in the development of the nonnormal filter is used to estimate spatial patterns and cannot force these patterns to be associated with a linear signal if they are not so in the unfiltered data.

Our results, as well as all of our previously published results (e.g., PS95, Penland and Magorian 1993), are also consistent with Kondrashov, Kravtsov, Robertson and Ghil (2005), who show that accounting for quadratic nonlinearities has an advantage in predicting extreme warm events as manifested by Niño 3 (5°S–5°N, 150°–90°W) SST anomalies, but otherwise is mostly indistinguishable from the linear model. Prediction of extremes is an important issue, and faster-than-seasonally varying nonlinearities (Kondrashov et al. 2005 used monthly rather than seasonal data) undoubtedly play an important role in maintaining extreme warm events. However, neither our studies nor theirs show any indication that nonlinear dynamics dominate the evolution of tropical SST anomalies except during the development of these warm extremes.

The importance of nonlinearities as a source of stochastic forcing must not be understimated. It has been stated by several of my colleagues that the term "noise" is unfortunate in that it implies processes to be filtered out if possible. In fact, these processes are crucial to the maintenance of those systems for which the linear approximation is valid, and the linear fluctuation-dissipation relation should be used as a guide for further study rather than an end in itself. After all, the linear approximation itself is best used as a diagnostic tool to investigate the strength of nonlinearities, thereby narrowing the possible physical mechanisms responsible for important weather and climate phenomena.

The time is ripe for research into the Central Limit Theorem properties of specific nonlinear processes. Are there timescales on which synchronized chaotic systems (e.g., Duane and Tribbia 2001; 2007) act as stochastic forcing and, if so, what are those timescales? We have already stated that LIM applied to dynamical models should reproduce the results of LIM applied to data. Conversely, it is time to apply nonlinear data analysis (e.g., Tsonis and Elsner 1996; Tsonis et al. 2006; Tsonis 2007) to realistic linear stochastic models and compare those results with those obtained from applying the analysis to data. If they are different, is there a timescale on which the results converge? Whether or not there is such a timescale, a careful analysis of the differences is likely to yield valuable information about the nature of the physical system.

Acknowledgements

The author is pleased to acknowledge travel support from NOAA's Environmental Research Laboratory, Physical Sciences Division. The discussion of the L63 spectra resulted from the very welcome comments of a referee, and I am grateful to Dr. P. D. Sardeshmukh for valuable discussions.

References

Archimedes, ed. Heiberg, vol. II, p. 244 (Arenarius I. 4-7) *The Works of Archimedes*, ed. Heath, pp. 221-222.

Aristarchus of Samos (310BCE-230BCE) Personal Communications to Archimedes. See Archimedes.

Battisti, D., and A. Hirst, (1989) Interannual Variability in a Tropical Atmosphere–Ocean Model: Influence of the Basic State, Ocean Geometry and Nonlinearity. J. Atmos. Sci. 46, 1687-1712.

Beck, C., and Roepstorff, G. (1987) From dynamical systems to the Langevin equation. Physica, 145A, 1-14.

Blanke, B., Neelin, J.D., Gutzler, D. (1997) Estimating the Effect of Stochastic Wind Stress Forcing on ENSO Irregularity. J. Climate 10, 1473-1486.

Cugliandolo, L. F., Dean, D.S., and Kurchan, J. (1997) Fluctuation-dissipation theorems and entropy production in relaxational systems. Physical Review Letters 79, 2168-2171.

Dorfman, J.R. (1999) Introduction to Chaos in Nonequilibrium Statistical Mechanics. Cambridge University Press, Cambridge.

Duane, G.S., and Tribbia, J.J. (2001) Synchronized Chaos in Geophysical Fluid Dynamics. Phys. Rev. Lett., 86, 4298-4301.

Duane, G.S., and Tribbia, J.J. (2007) Deterministic and stochastic parameter adaptation in the "Jungian" synchronization-based approach to data assimilation. *20 Years of Nonlinear Dynamics in Geosciences*, Springer, this volume.

Farrell, B., (1988) Optimal excitation of neutral Rossby waves. J. Atmos. Sci. 45, 163–172.

Garcia, A.L., Malek Mansour, M., Lie G., and Clementi, E. (1987) Numerical Integration of the Fluctuating Hydrodynamic Equations. J. Stat. Phys. 47, 209-228.

Hasselmann, K. (1988:) PIPs and POPs—The reduction of complex A general formalism for the reduction of dynamical systems in terms of Principal Interaction Patterns and Principal Oscillation Patterns. J. Geophys. Res. 93, 11015–11020.

Horsthemke W., and Lefever, R. (1984) *Noise-Induced Transitions: Theory and Applications in Physics, Chemistry and Biology*. Springer-Verlag, 318 pp.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. J. Educ. Psych. 24, 417-520.

Jarzynski, C. (1995) Thermalization of a Brownian particle via coupling to low-dimensional chaos. Physical Review Letters 74, 2937-2940.

Jin, F.-F., and Neelin, J.D. (1993*a*) Modes of Interannual Tropical Ocean–Atmosphere Interaction—a Unified View. Part I: Numerical Results. J. Atmos. Sci. 50, 3477-3503.

Jin, F.-F., and Neelin, J.D. (1993*b*) Modes of Interannual Tropical Ocean–Atmosphere Interaction—a Unified View. Part III: Analytical results in fully coupled cases. J. Atmos. Sci. 50, 3523-3540.

Jin, F.-F., Neelin, J.D., Ghil, M. (1994) El Niño on the Devil's Staircase: annual subharmonic steps to chaos. Science 264, 70-72.

Khasminskii, R.Z. (1966) A limit theorem for solutions of differential equations with random right-hand sides. Theory Prob. Applications 11, 390-406.

Kifer, Y. (2001) Averaging and climate models. *Stochastic Climate Models*, ed. P. Imkeller and J.-S. von Storch, Progress in Probability, 49, Birkhaueser, Basel.

Kleeman, R., and Moore, A.M. (1997) A Theory for the Limitation of ENSO Predictability Due to Stochastic Atmospheric Transients. J. Atmos. Sci. 54, 753-767.

Kolmogorov, A. (1933) *Grundbegriffe der Warscheinlischkeitrechnung*. Springer, Berlin.

Kloeden P. E., and Platen, E. (1992) Numerical Solution of Stochastic Differential Equations. Springer-Verlag, 632 pp.

Lorenz, E. (1963) Deterministic non-periodic flow. J. Atmos. Sci. 20, 130-141.

Lorenz, E. (1995) Predictability: a problem partly solved. *Seminar Proceedings "Predictability,"* European Center for Medium-Range Weather Forecasts, 1-18.

Majda A. J., Timofeyev, I., and Vanden Eijnden, E. (1999) Models for stochastic climate prediction. Proc. Natl. Acad. Sci. USA 96, 14687–14691.

Mo, K. C., and Häkkinen,S. (2001) Interannual Variability in the Tropical Atlantic and Linkages to the Pacific. J. Climate 14, 2740–2762.

Moore, A. M., and Kleeman, R. (1996) The dynamics of error growth and predictability in a coupled model of ENSO. Quart. J. Roy. Meteor. Soc. 122, 1405–1446.

Newman, M., Sardeshmukh, P. D., and Penland, C. (1997) Stochastic forcing of the winter time extratropical flow. J. Atmos. Sci. 54, 435-455.

Neelin, J.D., and Jin, F.-F. (1993) Modes of Interannual Tropical Ocean–Atmosphere Interaction—a Unified View. Part II: Analytical results in the weak coupling limit. J. Atmos. Sci. 50, 3504-3522.

Papanicolaou, G.C., and Kohler, W. (1974) Asymptotic theory of mixing stochastic ordinary differential equations. Comm. Pure Appl. Math. 27, 641-668.

Penland, C., and Magorian, T. (1993) Prediction of Niño 3 Sea-surface Temperature Anomalies using Linear Inverse Modeling. J. Climate 6, 1067-1075.

Penland, C., and Ghil, M. (1993) Forecasting Northern Hemisphere 700mb Geopotential Height Anomalies using Empirical Normal Modes. Monthly Weather Review 121, 2355-2372.

Penland, C., And Matrosova, L. (1994) A balance condition for stochastic numerical models with application to the El Niño - Southern Oscillation. J. Climate 7, 1352-1372.

Penland, C., and Sardeshmukh, P. D. (1995*a*) Error and sensitivity analysis of geophysical systems. J. Climate 8, 1988-1998.

Penland, C., and Sardeshmukh,P. D. (1995*b)* The optimal growth of tropical sea surface temperature anomalies. J. Climate 8, 1999-2024.

Penland, C. (1996) A stochastic model of Indo-Pacific sea surface temperature anomalies. Physica D 98, 534–558.

Penland, C., and L. Matrosova, L. (1998) Prediction of tropical Atlantic sea surface temperatures using Linear Inverse Modeling J. Climate, 11, 483-496.

Penland, C., and Matrosova, L. (2006) Studies of El Nino and interdecadal variability in tropical sea surface temperatures using a nonnormal filter. J. Climate 19, 5796-5815.

Ptolemy (148) *Almagest*, Canopus.

Rödenbeck, C., Beck, C., and Kantz, H, (2001) Dynamical systems with timescale separation. *Stochastic Climate Models*, ed. P. Imkeller and J.-S. von Storch, Progress in Probability, 49, Birkhaueser, Basel.

Sardeshmukh, P. D., Penland, C., and Newman, M. (2001) Rossby waves in a fluctuating medium. *Stochastic Climate Models*, ed. P. Imkeller and J.-S. von Storch, Progress in Probability, 49, Birkhaueser, Basel.

Sun, D.-Z., Zhang, T., and Shin, S.-I., (2004) The effect of subtropical cooling on the amplitude of ENSO: A numerical study. J. Climate 17, 3786-3798.

Sun, D.-Z. (2007) The role of ENSO in regulating its background state. *20 Years of Nonlinear Dynamics in Geosciences*, Springer, this volume.

Thompson, C.J., and Battisti, D.S., (2000) A linear stochastic model of ENSO. Part I: Model Development. J. Climate 13, 2818-2832.

Thompson, C.J., and Battisti, D.S., (2001) A linear stochastic model of ENSO. Part II: Analysis. J. Climate 14, 445-466.

Tsonis, A.A., and Elsner, J.B. (1996) Mapping the channels of communication between the tropics and higher latitudes in the atmosphere. Physica D 92, 237-244.

Tsonis, A.A., Swanson, K.L., and Roebber, P.J. (2006) What do networks have to do with climate? Bulletin of the American Meteorological Society 87, 585-595.

Tsonis, A.A. (2007) What do networks have to do with climate? *20 Years of Nonlinear Dynamics in Geosciences*, Springer, this volume.

Tziperman, E., Stone, L., Cane, M.A., Jarosh, H. (1994) El Niño chaos: Overlapping residences between the seasonal cycle and the Pacific ocean-atmosphere. Science 264, 72-74.

Von Storch, H., Bruns, T.,Fischer-Bruns, I., and Hasselmann, K. (1988) Principal Oscillation Pattern analysis of the 30–60 day oscillation in a GCM equatorial troposphere. J. Geophys. Res. 93, 11021–11036.

Wong, E., and Zakai, M. (1965) On the convergence of ordinary integrals to stochastic integrals. Annals of Mathematical Statistics 36, 1560-1564.

Woodruff, S., Lubker S., Wolter K., Worley S., and J. Elms, 1993: Comprehensive Ocean–Atmosphere Data Set (COADS) Release 1a: 1980–92. *Earth Sys. Monitor* 4, 1–8.

28 Reducing Forecast Uncertainty to Understand Atmospheric Flow Transitions

Paul J. Roebber[1] and Anastasios A. Tsonis[2]

[1] University of Wisconsin at Milwaukee, Atmospheric Science Group, Department of Mathematical Sciences, roebber@uwm.edu
[2] University of Wisconsin at Milwaukee, Atmospheric Science Group, Department of Mathematical Sciences, aatsonis@uwm.edu

Abstract. A key forecast challenge is to anticipate transitions in the atmospheric flow. Ensemble prediction has become an indispensable tool in weather forecasting, as a means of quantifying the relative likelihood of particular future states. One of the issues in ensemble prediction, however, is that regardless of the method, the prediction error does not map well to the underlying physics (in other words, error estimates do not project strongly onto physical structures). In this work, we show through a series of experiments evidence that a nonlinear approach will allow us to isolate and eliminate a deterministic portion of the error in an ensemble prediction. This, in combination with appropriate diagnostics, will first enable a more accurate prediction of transitions in the large-scale atmospheric flow and second, more cleanly relate the error to the underlying state, thereby gaining insight into the predictability of such transitions. The experimental basis is provided from results obtained from three increasingly sophisticated chaotic systems, which incorporate atmospheric transitions and hemispheric structure. Using neural networks to probe the deterministic component of forecast error, we show that the error recovery relates to the underlying type of flow and that it can be used to better-forecast transitions in the atmospheric flow using ensemble data. We discuss a diagnostic method that might be employed on more sophisticated atmospheric models such as the full primitive equation ensemble-modeling systems currently employed in operations.

1 Introduction

Since the mid-latitude atmosphere is characterized by variability, persistent flow patterns often lead to anomalous sensible weather (e.g., temperature, precipitation). Yet, one challenge in short to medium range weather prediction is anticipating transitions in the large-scale atmospheric flow. Hence, an improved understanding of the physics and predictability of such flow transitions is desirable.

An example of this is provided by a so-called Hovmöller plot of the middle atmospheric flow pattern (strictly, the height of the constant 500 hPa surface, averaged between 30 and 40°N) for a three month period in the spring of 2006 (Fig. 28.1). A persistent trough is evident over the western United States during the early spring,

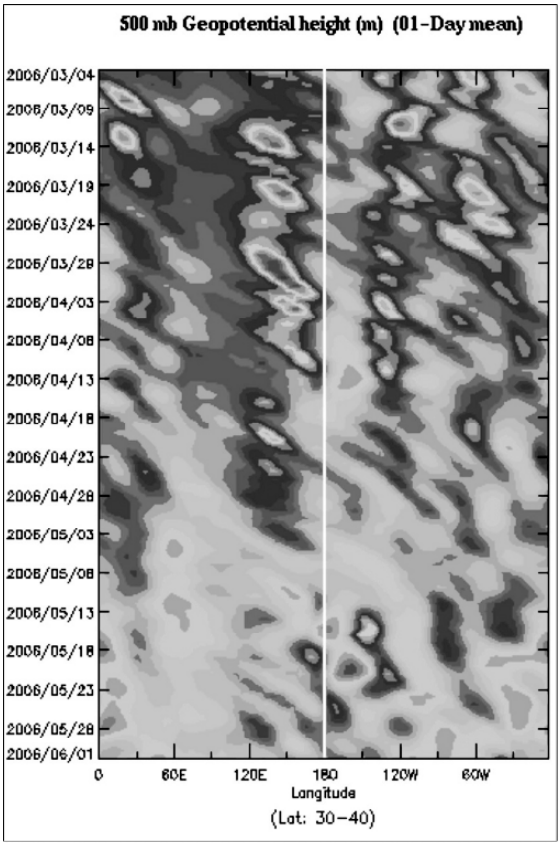


Fig. 28.1. Hovmöller plot of 500 hPa height for 30-40°N for March-May 2006. (A color version of this figure appears between pages 196 and 197).

which gives way to a ridge in May. This is reflected by the monthly temperatures, which are anomalously cool in March (Fig. 28.2) and warm in May (Fig. 28.3) in that region.

The fundamental problem of weather forecasting is to identify the range of possible meteorological scenarios that might evolve from a given initial state, and determine whether multiple solutions have high probability (indicating low confidence in an individual solution) or if a single evolution is the most likely (with corresponding high confidence). This probabilistic view is necessitated by the complexity of atmospheric dynamics (e.g. Lorenz 1963; see a general review in Kalnay 2003). Specifically, limits to deterministic predictability originate from two general sources: model error and initial condition error.

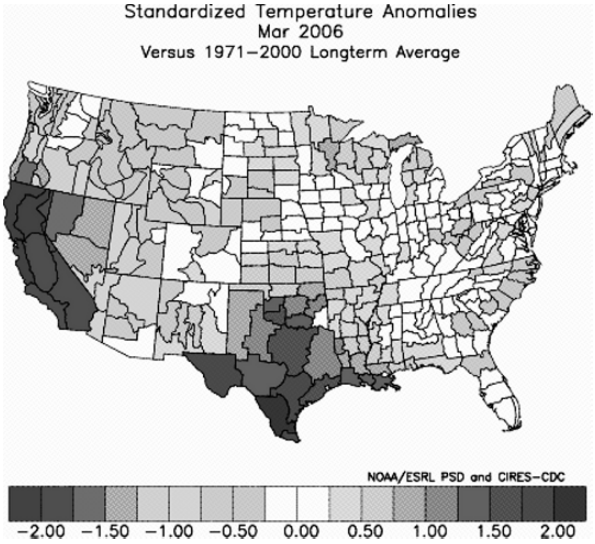


Fig. 28.2. Standardized Temperature anomalies for the continental United States for March 2006. (A color version of this figure appears between pages 196 and 197).

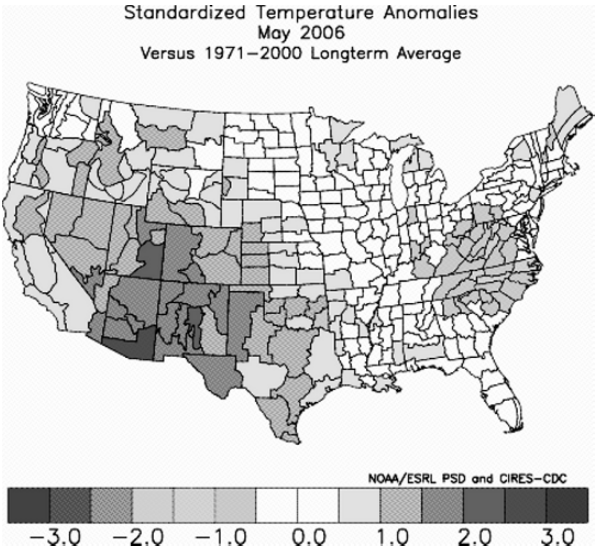


Fig. 28.3. Standardized Temperature anomalies for the continental United States for May 2006. (A color version of this figure appears between pages 196 and 197).

Model error arises because of imperfections in our characterizations of the laws of nature, arising either through parameterization of complex and/or poorly understood

physics (such as boundary layer and cloud microphysical processes) or an inability to resolve atmospheric processes smaller than a certain threshold (e.g., atmospheric convection with a 10 km grid point model), with resultant upscale error growth.

Initial condition error arises because of the finite spatial availability of observed data (including some variables that are not observed at all), missing data, inaccuracies in the data and imperfect analysis techniques. All of these errors, even with a perfect model, will grow nonlinearly over time, eventually swamping the forecast signal (Lorenz 1963; 1965; 1968).

The rate of error growth and hence the lead time at which predictability is lost depends on the stability of the evolving flow (Lorenz 1965), which in addition is affected by factors such as the large scale flow pattern, season, and geographical domain (Lorenz 1984; 1990). Ensemble forecast systems have been developed as a means to quantify forecast uncertainty, using a variety of methods to simulate analysis and model uncertainties. In the middle latitudes, numerical weather prediction models are often considered to be good enough that to a first approximation, forecast error growth at medium range can be attributed primarily to the process of growing instabilities in a chaotic system that result from initial condition uncertainties. This "perfect model" assumption is more prone to failure, however, for short-range forecasts where small-scale features may depend crucially on the physics (e.g., convection) or where substantial flow transitions are occurring. For this reason, we also consider model error in this work (section 2.3).

A deeper understanding of the physical evolution of error structures can potentially reveal much concerning both the natural dynamics of the atmosphere and model deficiencies (and hence, lead to a better understanding of nature). Currently, there are four primary methods for assessing analysis uncertainty: breeding [until recently, used at the US National Centers for Environmental Prediction (NCEP)], singular vectors (used at the European Centre for Medium-Range Weather Forecasts), Kalman filters (currently a topic of active research and now implemented at NCEP) and Monte Carlo methods (used at the Canadian Meteorological Centre).

Although the technical details of each of these methods are distinct (see Kalnay 2003 for an overview), they each rely on a single concept: initial analysis uncertainties are simulated by adding small perturbations (within the limits of the uncertainty of the analysis) to the unperturbed (control) analysis. It is important to recognize that these techniques are intended to simulate analysis uncertainty, not necessarily to reveal intrinsic dynamics. Tribbia and Baumhefner (2004) argue that these uncertainties seed the baroclinically active region of the spectrum and organize over time on the synoptic scales, extracting energy from the larger-scale flow rather than the smaller scales.

In the breeding method, analysis errors are considered to be filtered forecast errors. As such, much of the analysis error may represent phase error, a typical problem in data sparse regions. The singular vector method attempts to directly calculate forecast error, under the assumption that the components of the error that grow most rapidly will dominate. In practice, the actual error may project more or less strongly onto these growing patterns and experience with this method suggests that these patterns do not "map" directly onto locations that are consistent with synoptic experience as to the origins of forecast uncertainties (e.g., the position of an upper level

jet streak). The Kalman filter method uses estimation theory to more accurately express the breeding method's fundamental hypothesis that analysis errors are filtered forecast errors. As such, this method likely preferentially identifies phase errors, which are consistent with the underlying attractor. The Monte Carlo method perturbs all available observations simultaneously with random numbers of realistic amplitude (with respect to error statistics). Hence, no direct dynamics can be inferred from this method.

It is well known that, given a nonlinear system, the evolution of small fluctuations is given by the linearized equations:

$$\dot{\vec{x}}' = J\vec{x}' \tag{1}$$

where J is the Jacobian. These fluctuations can be thought of as an unbiased measurement error, e_{nb}, and/or a biased error arising through the data assimilation process, e_b, which contribute to the forecast error (this latter case is also a form of model error). Since the Jacobian implies deterministic operations, some recoverable dynamics are included in the error. This conclusion also holds for the more general, nonlinear problem.

Conceptually, the idea is as follows. As discussed by Palmer (1995) (see his Fig. 2.1), in the early part of the forecast (2-3 days), error growth is governed by linear dynamics. The probability density function (PDF) evolves from an m-dimensional ball to an ellipsoid, with the major axis corresponding to the dominant instability. The PDF than progresses to the weakly nonlinear stage (up until about day 7), wherein the ellipsoidal shape is deformed and there is a change in the major axis direction. Following this, the PDF growth is strongly nonlinear and there may be little correspondence between the major axis obtained in the linear stage and the dominant directions of growth. As Palmer (1995) notes, however, despite this nonlinear growth, predictability has not necessarily been lost. We demonstrate in section 2, using three chaotic systems of increasing complexity, that a nonlinear method can capitalize on *recovering* part of the forecast error provided by any of the above ensemble approaches and then *relate that error recovery to the underlying state of the flow*. Specifically, the technique allows one to obtain an improved measure of the forecast uncertainty which can be used in concert with diagnostics to relate the evolution of the error to the underlying flow.

The use of simple models to gain insights into complex problems has a rich history in meteorological research (recent examples are the investigations of the dynamics of model error by Nicolis 2003; 2004). The low order modeling approach is undertaken on three models in this Chapter (the recursive logistic equation, a low-dimensional climate model with identifiable flow patterns and a low-dimensional, but non-local atmospheric model) with the expectation that it may reveal general principles, the details of which will need to be determined by resorting to studies using more complex models. In section 3, we outline a possible extension of this technique to full atmospheric models, with the objective of improving understanding of the predictability of transitions in the large-scale atmospheric flow.

2 Examples

2.1 The logistic equation

First, we consider a very simple mathematical system, which however can be extremely sensitive to initial conditions. It is the logistic equation:

$$X_{n+1} = \mu X_n (1-X_n) \tag{2}$$

where μ is a parameter. In this very simple discrete system, future values are obtained by successive iterations. The system exhibits a variety of dynamics for different values of μ. For μ=4 the system is strongly chaotic with a corresponding Lyapunov exponent equal to ln(2). This means that if the initial condition is specified with an error of less than 1%, the forecast is no better than "climate" after approximately six iterations.

For the experiments with the logistic model, the operational data assimilation cycle was mimicked: an analysis was constructed, consisting of a weighted average of observations (incorporating an unbiased error $e_{nb,o}$) and a first guess (obtained from a previous short-term forecast from the model). This analysis was used as the basis for the next forecast, and the short-term forecast from that run was then cycled back into the succeeding analysis. For all our tests, the set of forecasts were extended from two to four iterations into the future, a time scale for which substantial error growth occurs in the logistic equation with these error characteristics (Fig. 28.4).

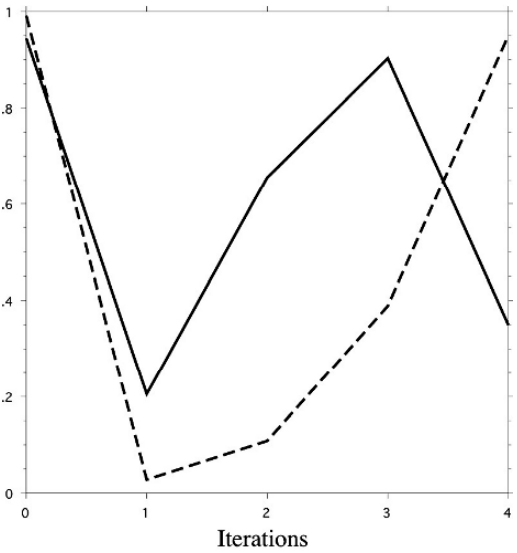


Fig. 28.4. Logistic equation "control" and "perturbed" initial state runs, through 4 iterations.

Error recovery was investigated using neural networks. Neural networks are nonlinear tools that are often used to model nonlinear deterministic dynamics when the actual equations describing the dynamics are not available. In its simplest form the architecture of a neural network involves a set of inputs *p[i]* and an output *Q*. Using a training sample the inputs are assigned optimum weights, *w[i]*, and then the inner product, $P=\Sigma p[i]w[i]$, is estimated. The final output *Q* is then obtained by passing *P* through a nonlinear function *f(x)*. Such a network is called a two-layer network since it involves a layer for the inputs and a layer for the output. Neural networks can become more complex by adding hidden layers or by including more than one output. In general, with the help of known outputs (from the training sample) the network, initially set to a random state, modifies its structure (changes the weights) in such a way as to improve its performance. If the network architecture is rich enough, this procedure leads the network to a state where inputs are successfully mapped into outputs for all chosen training (input-output) pairs. Essentially, the network describes a nonlinear model fit of the error. Further information on neural networks can be found in Tsonis (1992), Roebber, Bruening, Schultz and Cortinas (2003), Marzban (2003) and references therein.

The procedure to train and evaluate the network was as follows (note: this procedure is followed for all the networks discussed herein, except where otherwise noted). Data were evenly split into training and testing samples. Training was accomplished using 10% of the training data for cross-validation (to avoid overfitting). The error in the cross-validation dataset was monitored and training was stopped when this error began to increase. The cost function used in the training was the mean-square error. The deepest local minimum in the training over several reinitializations was taken to be the global minimum.

A generalized feed-forward network with one hidden layer (and 70 nodes) was used for the error recovery experiments. The optimal number of nodes was determined using cross-validation. The activation functions were hyperbolic tangent, while a momentum learning rule was used for training (to speed up and stabilize convergence relative to straight gradient descent). Nine inputs were used: the initial analysis, the observations (including error) at the initial time and the lagged forecasts at 2, 3 and 4 time steps validating at the verification time.

Correlation coefficients between "forecasts" from the logistic equation and the verification (using the unperturbed initial conditions) show that, at 2 time steps, substantial predictability exists (r=0.80; Table 28.1). Despite this high predictability, the forecasts can be improved still further (~17% relative to the 2 iteration forecast, based on the increase in accounted variance) with the trained neural network. This exceeds by 8% the predictability obtained with a linear regression whose inputs are the set of lagged forecasts with regression coefficients optimized using the training dataset (not shown). Hence, in this highly chaotic system, it is still possible to recover a significant fraction of the total forecast error using the nonlinear technique (~48%, based on the amount of accounted variance relative to the maximum possible improvement over the 2 iteration deterministic forecast).

2.2 The Lorenz (1984) climate model

While the above is an example of a general chaotic system on which one could test our hypothesis, it has little relationship to atmospheric phenomena. As such, it leaves unanswered the question as to whether this approach can be applied to large-scale atmospheric flows. For this reason we consider another example, which represents a low-order general circulation model (Lorenz 1984; 1990). It is also highly chaotic and is described by the following equations:

$$\begin{aligned} \frac{dX}{dt} &= -Y^2 - Z^2 - aX + aF \\ \frac{dY}{dt} &= XY - bXZ - Y + G \\ \frac{dZ}{dt} &= bXY + XZ - Z \end{aligned} \tag{3}$$

These equations can be derived from the mean and perturbation quasi-geostrophic potential vorticity equations, using a truncated Fourier series expansion of the geostrophic stream-function (Holton 2004). The independent variable t represents time, while X, Y, and Z represent the meridional temperature gradient (or equivalently from thermal wind balance, the strength of the zonal flow), and the amplitudes of the cosine and sine phases of a chain of superposed large scale eddies, respectively. The term F represents the meridional gradient of diabatic heating, the value to which X would be driven in the absence of coupling between the westerly current and the eddies. The term G is the asymmetric thermal forcing, representing the longitudinal heating contrast between land and sea, and is the value to which Y would be driven in the absence of coupling. The coupling is represented by the terms XY, XZ, and $-Y^2$-Z^2, and results in amplification of the eddies at the expense of the westerly current.

With F=8 and G=1 (representing perpetual winter conditions; see Lorenz 1984; 1990; Roebber 1995), the above system exhibits two very distinct flow patterns (Fig. 28.5). The first, which consists of low amplitude fluctuations, represents a steady, zonal jet. The second, consisting of high amplitude fluctuations, corresponds to alternating strong and weak jets. Another regime is that of a blocking event, but its frequency of occurrence for these parameters is very small and highly transient. For our analysis, this pattern will not be considered.

Predictability in this system is strongly dependent on initial conditions (Fig. 28.6). In this example, a forecast starting with a small error in the initial state results in an error that is quite large by three time units (representing 120 time steps), the result of a forecast transition from a high amplitude to a low amplitude flow that did not verify. In other instances, of course, pattern transitions will be correctly forecast and the errors will be less.

Experiments with the Lorenz model were designed as follows. The model was run without interruption for an extended period (1164 time units) and the data

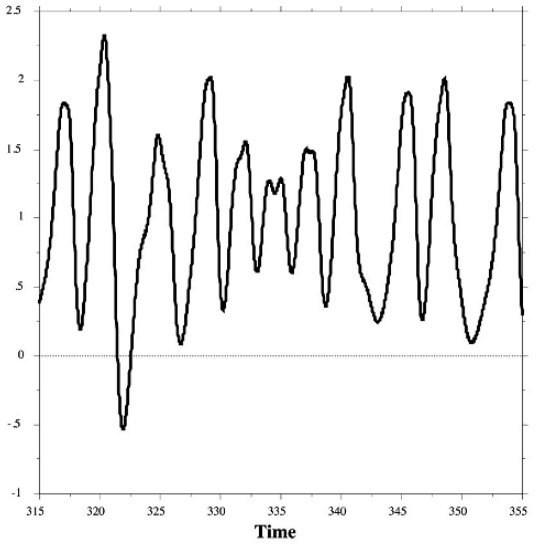


Fig. 28.5. Lorenz (1984) climate model integration for a 40 time unit period. The value of X is shown. High amplitude oscillations are apparent from time=315 through 330, followed by low amplitude oscillations through time=340, and back again to high.

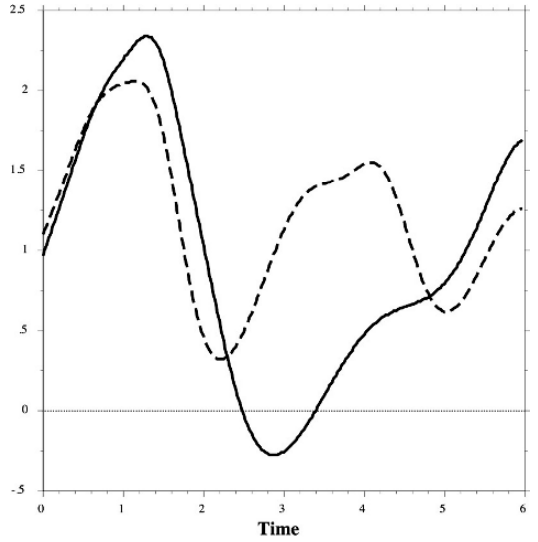


Fig. 28.6. Lorenz (1984) climate model integrations for a 6 unit time period, starting from an initial state and a second integration from a slightly perturbed initial state. The value of X, the meridional temperature gradient from the model, is shown.

recorded (base run). The model was then rerun 194 times out to 6 time units (6x194=1164), with the initial conditions obtained from the corresponding values of the base run with a small error superimposed (mean absolute error for the initial conditions of the resulting 194 runs is 0.08 in *X*, *Y* and *Z*). Of these 194 runs, 98 (96) were high-amplitude (low-amplitude) flows.

Table 28.1 shows that for the low-amplitude pattern at 2 time units, predictability is relatively high (r=0.77), but that it is still possible to recover a substantial portion of the error using neural networks sorted according to the flow (note that the network design is the same as that used for the logistic equation, except with 7 inputs representing the initial analysis and observed values for *X*, *Y* and *Z* at lead times of either 2 or 4 units, plus the forecast value for *X* at the verification time).

In contrast, predictability is much higher for the high amplitude pattern and little additional predictability can be achieved. By 4 time units, these differences are accentuated (Fig. 28.7), with limited correspondence between the forecast and observed states for the low amplitude pattern, but substantial error recovery using a neural network with the initial analysis, observations (including error) at the initial time and the forecast as inputs (Table 28.1). Figure 28.8 is similar to Fig. 28.7, but shows the corrected forecasts using the neural network. Comparison between Figs. 28.7 and 28.8 clearly shows that the error variance of the high amplitude pattern remains approximately the same following non-linear correction, but that the error variance of the low amplitude cases decreases significantly.

This finding indicates, as suggested earlier, that the error recovery is flow dependent, a result that is consistent with previous research that documents flow dependent predictability in the observed atmosphere (e.g. Nutter, Mullen and Baumhefner 1998). This suggests that it should be possible to use this differential error recovery to map the error to the underlying physical characteristics of the zonal jets. Since the varying flow is dictated by the physics of the system, this would indicate that error recovery could be used to understand the physics. Here, the low amplitude state in the model occurs for relatively large available potential energy and small kinetic energy, such that a flow transition also reflects a substantial transition in the energy budget. Importantly, this result also holds for pattern transitions [in which the initial flow is a low (high) amplitude flow which transitions to a high (low) amplitude flow by forecast verification time], such that an additional 9-12% of the variance can be accounted for at 2-4 time units in these cases compared to those in which a particular pattern is maintained throughout the forecast interval (Table 28.1).

These results suggest that differential error recovery also can be used within an ensemble forecast context to forecast flow transitions. To test this idea, the Lorenz model was used to generate ensemble forecasts for a set of events in which both patterns verify. The reduction of the variance in the model ensemble forecasts (as opposed to the error, which is unknown in forecast mode) using the previously trained neural network (with a median ensemble size of 8 members) is used to classify the forecast as representing either the low or high amplitude pattern (Table 28.2). The classification is performed using a simple probabilistic model assuming that the error recovery is normally distributed (a hypothesis that cannot be rejected at the 1% level of significance, based upon the Kolomogorov-Smirnov test of Normality).

Experiment	*Deterministic Forecast*	*Neural Network*
Logistic – 2 iterations	0.80	0.90
Lorenz – low – 2 time units	0.77	0.87
Lorenz – low – 4 time units	0.35	0.67
Lorenz – high – 2 time units	0.95	0.97
Lorenz – high – 4 time units	0.78	0.84
Lorenz – trans – 2 time units	0.83	0.94
Lorenz – trans – 4 time units	0.61	0.83

Table 28.1. Correlation coefficient between forecasts and "truth" for the logistic and Lorenz (1984) models. The Lorenz model verifications are stratified according to the low and high amplitude flow patterns, and pattern transitions from low-to-high and high-to-low.

Reduction in Standard Deviation of X (meridional thermal gradient)

Forecast range	*Final High Amplitude*	*Final Low Amplitude*
2 time unit ensemble	0.0607	0.1089
4 time unit ensemble	0.0675	0.2608

Ensemble Pattern Diagnosis – Contingency Table (# of cases)

Pattern at Final Time	*Diagnosed Final Pattern*	
	Low Amp	*High Amp*
Low Amp	73	22
High Amp	9	83

Measures-based Verification of Network Pattern Diagnosis

Pattern at Final Time	*Probability of detection*	*False Alarm Rate*	*Bias*	*True Skill Statistic*
Low Amp	0.77	0.11	0.86	0.67
High Amp	0.90	0.21	1.14	0.67

Table 28.2. Error recovery and flow pattern diagnosis in the Lorenz (1984) model. Shown are the reduction in the standard deviation of the ensemble forecasts using the neural network for the two patterns at two time ranges; the contingency table of pattern diagnostics and associated verification statistics.

Standard measures-oriented skill verifications show that this technique is highly successful in the context of this two pattern model (note that the true skill statistic for the ensemble mean forecast is a considerably lower 0.32). These experiments indicate that the theoretical possibility of partial error recovery can be realized even in highly nonlinear systems, and that it may be possible to apply techniques of this kind to the output of ensemble systems to better predict and understand atmospheric flow transitions.

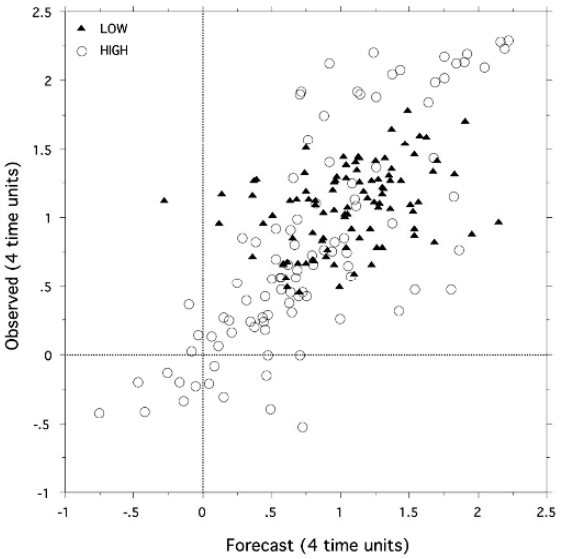


Fig. 28.7. Scatterplot of Lorenz (1984) model forecasts (4 time units from initial state) and "observed" values of X (meridional temperature gradient)for the two flow patterns.

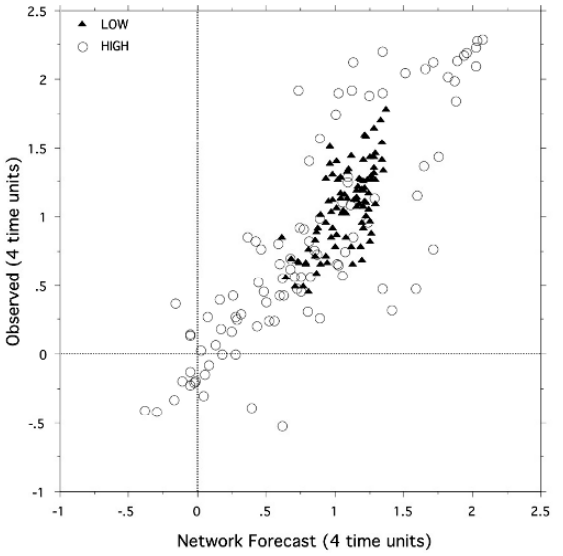


Fig. 28.8. Scatterplot of neural network corrected forecasts (4 time units from initial state) and "observed" values of X (meridional temperature gradient) for the two flow patterns. Note the reduction of variance in low amplitude cases relative to Fig. 28.7.

2.3 The Lorenz (1995) hemispheric model

So far, we have investigated an approach in which a nonlinear method is used to recover the deterministic component of the forecast error. Using two low-order nonlinear models we show that not only this is possible, but that this error recovery reduces the forecast error enough to allow us to map the remaining error to the underlying state and to improve the prediction of transitions. Success in correcting other types of forecasts with neural networks have been obtained in realistic forecast settings (e.g., Hall, Brooks and Doswell III 1999; Roebber et al. 2003; Marzban 2003), so there is reason to suppose that this approach might be useful with actual atmospheric data.

Some additional aspects must be considered, however, before it can be applied to real ensemble forecast systems. First, the simple ensemble system studied in this preliminary analysis effectively represents a single point; hence, this local error correction method must be generalized to account for the effects of error propagation in real systems. Second, this study has been conducted under the constraints of the perfect model assumption. As noted in section 1, model error imposes an important limit on forecast accuracy.

In order to address these aspects, we turn to experiments with another simple (but nonlocal) chaotic atmospheric model (Lorenz 1995). This model, slightly modified to allow downstream progression, contains 36 variables X_1, …, X_{36} and is governed by 36 equations of the form:

$$\frac{dX_k}{dt} = -X_{k-2}X_{k-1} + X_{k-1}X_{k+1} - X_k - c(X_{k+1} - X_{k-1}) + F \qquad (4)$$

where k ranges from 1 to 36 and the constants F and c are independent of k. The X_k can be thought of as values of 500 hPa height in 36 sectors of a latitude circle. This structure, which conserves total energy, simulates external forcing, internal dissipation and advection. With the external forcing F equal to 8 in this system, the error doubling time is approximately 2.1 days, consistent with that of global circulation models (Lorenz 1995). Hovmöller type plots of the output from this model indicate that it produces a reasonable simulation of observed atmospheric evolutions, in the sense that the flow can retrograde, remain quasi-stationary and progress, while changing amplitudes (Fig. 28.9).

A perfect model experiment is established as follows. A simulation comprising 30 years of model time is produced using Eq. 4 to represent the daily evolution of the actual atmosphere. This 30-year time frame is imposed to simulate the availability of actual ensemble forecast data and test feasibility. Then, daily 10-member ensemble forecasts out to 12 days range are produced for the entire 30 years using Eq. 4 with the "actual" data plus a small unbiased, random error as initial conditions (the median absolute initial condition error is 2.5% for individual members). Because of the symmetry of the model, each longitude exhibits statistically similar behavior. Hence, for the ensuing training and testing, we focus on a single longitude (k=18).

At 10 day range, individual ensemble members correlate with the "observations" with r=0.56, while the 10-member ensemble mean improves to r=0.79, with a

tendency for the ensemble mean to underpredict the extreme events. Further, forecast variability within the ensemble positively correlates with ensemble mean forecast error (r=0.56), a result consistent with theoretical expectations in a perfect model environment (unfortunately, the predictive value of this information by itself is limited, since the variability is unable to provide an indication of whether the errors will be positive or negative in a given case).

In order to extract temporal information contained in the data, an Elman-type neural network, a form of the multilayer perceptron (MLP) extended with context

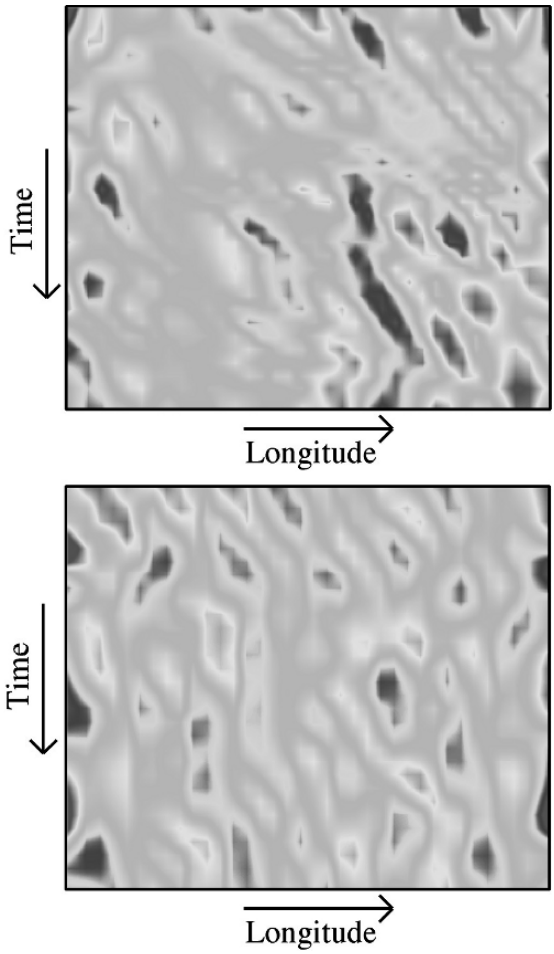


Fig. 28.9. Observed (top) and Lorenz (1995) modeled (bottom) 500 hPa flow for two randomly selected 30 day periods. Dark (light) are negative (positive) height anomalies.

units (processing elements that "remember" past activity, in this case, past inputs), was trained to correct the forecasts (the target was the "analysis" rather than the actual state, which in practice is not known). Inputs in the training were the 10 day, 11 day and 12 day ensemble mean forecasts validating for the same date, and the standard deviation of the 10 day ensemble forecast derived from the 8 members of the ensemble. The sequence of lagged forecasts provides information concerning trends in the forecast (likely representing the impact of upstream error propagation), while the ensemble forecast variance provides a measure of the uncertainty, as previously noted. The time-dependent network design facilitates the extraction of trends in the forecast inputs. The network contained one hidden layer with 140 nodes, with a linear integrator in the context units.

The output from this network provides a 2.5% improvement in accounted variance relative to the ensemble mean. For those cases with large ensemble variance (as defined by ensemble standard deviations above the 75th percentile), however, the network improvement is 6.7% relative to the ensemble mean. Such variability in medium range forecasts is often associated with flow transitions, indicating that the technique may be most useful in these important situations. Indeed, in the cases where the network error was low and the ensemble mean error was not (defined by the 25th percentile of the error), the flow was characterized by moderate amplitude features with an upstream trough and a downstream ridge. In contrast, low error in the ensemble mean was preferentially associated with zonal flows.

We next ask, what is the impact of model error on our ability to recover the forecast signal? An imperfect model experiment is established by producing daily 10-member ensemble forecasts out to 12 days range as before. However, in Eq. 4, F is set to 8.5 to represent a deviation from the perfect model with F=8. This change increases the error doubling time (Lorenz 1995). Here, the spread-skill relationship is less good, with r=0.35 (comparable to NCEP ensemble model results at medium range as reported by Whitaker and Loughe 1998). At 10 day range, individual ensemble members correlate with the observations with r=0.37, while the 10-member ensemble mean improves to r=0.53. An Elman-type network is trained as before with the same inputs. In this instance, the network output provides an 11.6% improvement in accounted variance relative to the ensemble (23.8% when considering those cases with large ensemble variance). Hence, in the realistic setting of limited training data and an imperfect model, the nonlinear correction technique appears increasingly valuable as far as reduction of error in the ensemble mean.

An important extension to this is in the estimation of forecast uncertainty. The ensemble system itself provides a measure of forecast uncertainty through the spread of the individual members. What benefit does the neural network technique provide in this context? In order to address this, the trained network was used to correct the individual members of the ensemble (inputs were as above, except that each individual member was substituted for the ensemble mean for each of the 10 members in turn, and the network rerun). This results in a new ensemble, consisting of the 10 corrected individual members, which is (unsurprisingly) statistically similar to the original, corrected ensemble (correlation between the two ensemble means is 0.998). What is gained, however, is an improved measure of forecast uncertainty. The new

ensemble shows a gain in the spread-skill relationship of 5.5% relative to the raw ensemble standard deviation.

We have shown that there is both theoretical and experimental evidence that a nonlinear approach will allow us to isolate and eliminate a deterministic portion of the error in an ensemble prediction. We have shown that this is true even in the context of non-local errors, and further that the approach is most effective where significant ensemble variability exists, such as associated with flow transitions, and where the models are imperfect. We have shown that an improved measure of forecast uncertainty is obtainable.

The remaining question is how to connect the findings from these preliminary experiments to the real atmosphere? In the next section, we discuss a diagnostic that might be used to relate the error obtained from these methods to the underlying state, thereby gaining insight into the predictability of such transitions.

3 A Future Diagnostic

We desire a method that will allow a direct examination of the physics of the flow-dependent error in a full physics global forecast model. In order to accomplish this, we will need an estimate of the spatial structure of the evolving forecast errors. One natural approach follows from section 2.3: employ "Hovmöller" networks, as in the case of the Lorenz (1995) hemispheric model. A measure of the flow (e.g., 500 hPa height anomalies) at a given longitude for a specified latitude band could be used as inputs to an Elman-type neural network. The output from the network would be the corrected forecast of the original ensemble (note that as before, the lagged ensemble mean forecasts and the variation of the individual members make up the input to the network to simplify training; the individual members of the ensemble are used for production of a corrected ensemble mean and spread). These corrected flow measures, which are quite distinct from the raw ensemble data, can then be examined, in combination with verification data, to determine the spatial structure of the corrected errors within specific flow patterns and their relation to the corresponding state (dynamics) of the flow. Unlike the statistically uniform Lorenz (1995) model, longitudinal variation in the "climate" of the real atmosphere can be significant. Hence, the correction procedure would be applied to a series of longitudes, within the mid-latitude band, using independently trained networks. This procedure could be extended as needed to multiple vertical levels and latitude bands. Whether such steps are necessary will need to be informed by actual experiments.

An appropriate methodology for studying the dynamics of large-scale flow transitions has been presented by Nielsen-Gammon and Lefevre (1996) and Evans and Black (2003). This method involves a partitioning of the local time tendency of quasi-geostrophic potential vorticity into advective and nonconservative forcing terms, and through piecewise inversion, obtaining the height-tendencies induced by that forcing. Since height anomalies are closely linked with circulation anomalies, this formalism will allow for a quantitative examination of the dynamics (specifically, linear baroclinic and barotropic growth mechanisms and nonlinear interactions; see Evans and Black 2003 for details) during flow transitions associated with

particular error structures. Individual and flow pattern composite analyses could be used to examine the predictability of these transitions.

As a simple example, consider results obtained from the Lorenz (1995) model experiments. Transitions from zonal to meridional flow are found to be associated with high ensemble forecast error. Reduction of the error using the neural network technique was found to be significant in these instances (23.8%, when considering an imperfect model setting). Given similar behaviors in a full physics ensemble, one could proceed as follows. First, the reduced error would be used to quantify the evolution of the error structures and to measure the forecast uncertainty during these transitions. Second, using the piecewise tendency diagnosis technique, one could investigate the dynamics of those flow transitions in which the corrected forecast uncertainty is high versus those cases in which it is not. What are the differences in the operative physics in these instances? How do the corrected error structures map onto the physics? Evans and Black (2003) found that for high amplitude Pacific circulation anomalies of both signs, onset was governed primarily by baroclinic growth, with secondary contributions from barotropic deformation. In the Atlantic, in contrast, barotropic growth was primary. We hypothesize that differences in baroclinic versus barotropic growth processes may have an important influence on predictability, but this remains to be demonstrated.

It is known that neural networks require relatively extensive data for effective training, in order to represent every part of the multidimensional input space and to protect against memorization. This is particularly true in the context of hemispheric or global models, where the dimensionality is necessarily large. In the experiments reported here, it was of course possible to generate as much data as needed for effective training. In order to evaluate the likelihood of obtaining satisfactory results using an existing ensemble dataset (e.g., the Climate Diagnostic Center Reforecast Ensemble, which includes data since 1979), however, we have deliberately restricted the generated data in our experiments with the Lorenz (1995) model. In doing so, we have established that it is possible to train a "Hovmöller" type of network within real data limitations.

There are two additional issues related to error structures that should be explored in the context of real data. First, what is the impact of the hemispheric structure of initial condition error, which exists owing to the relative paucity of observations over the oceans? Based on our preliminary experiments, which showed the greatest improvements where ensemble variance (and hence forecast error) was largest, it seems likely that the greatest corrections will be obtainable over these data sparse regions. Second, what is the impact of the temporal error correlation that also results from the geographic variation in observation density (the larger errors over the oceans will tend to persist until they are corrected by an observation)? Again, based on our preliminary analyses, these errors will likely be preferentially corrected; further, the temporal character of these errors will be accommodated by the network and data input design, which feature temporal aspects.

4 Summary

In this work, we have investigated a nonlinear approach that allows us to isolate and eliminate a deterministic portion of ensemble prediction error. Experiments with this technique are conducted using data from three increasingly sophisticated chaotic systems, starting with the logistic equation and continuing with the Lorenz (1984) climate model (which incorporates atmospheric transitions) and finally the Lorenz (1995) hemispheric model. Neural networks are used to probe the deterministic component of the forecast error in each of these systems. We show that the error recovery relates to the underlying type of flow and that it can be used to better-forecast transitions in the atmospheric flow using ensemble data. For these systems, we show that a more accurate prediction of transitions can be produced, and that using appropriate diagnostics, we can relate the error to the underlying state and thereby gain physical insight into their predictability. For more sophisticated atmospheric models such as the full primitive equation ensemble-modeling systems currently employed in operations, we discuss a diagnostic method that might be employed.

References

Evans, K.J. and Black, R.X. (2003) Piecewise tendency diagnosis of weather regime transitions. J. Atmos. Sci. 60, 1941-1959.

Hall, T., Brooks, H. E. and Doswell III, C. A. (1999) Precipitation forecasting using a neural network. Wea. Forecasting 14, 338–345.

Holton, J.R. (2004) *An introduction to dynamic meteorology*, 4th edition. Elsevier Academic Press, London.

Kalnay, E. (2003) *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, Cambridge.

Lorenz, E.N. (1963) Deterministic non-periodic flow. J. Atmos. Sci. 20, 130-141.

Lorenz, E.N. (1965) A study of the predictability of a 28-variable atmospheric model. Tellus 17, 321-333.

Lorenz, E.N. (1968) The predictability of a flow which possesses many scales of motion. Tellus 21, 289-307.

Lorenz, E.N. (1984) Irregularity: A fundamental property of the atmosphere. Tellus 36A, 98-110.

Lorenz, E.N. (1990) Can chaos and intransitivity lead to interannual variability? Tellus 42A, 378-389.

Lorenz, E.N. (1995) Predictability - A problem partly solved. In: *Seminar Proceedings of the ECMWF seminar on predictability*, Reading, 4-8 September 1995, Vol. I, pp. 1-18.

Marzban, C. (2003) Neural networks for postprocessing model output: ARPS. Mon. Wea. Rev. 131, 1103-1111.

Nicolis, C., (2003) Dynamics of model error: some generic features. J. Atmos. Sci. 60, 2208-2218.

Nicolis, C. (2004) Dynamics of model error: The role of unresolved scales revisited. J. Atmos. Sci. 14, 1740-1753.

Nielsen-Gammon, J., and Lefevre, R. (1996) Piecewise tendency diagnosis of dynamical processes governing the development of an upper-tropospheric mobile trough. J. Atmos. Sci. 53, 3120-3142.

Nutter, P.A., Mullen, S.L. and Baumhefner, D.P. (1998) The impact of initial condition uncertainty on numerical simulations of blocking. Mon. Wea. Rev. 126, 2482-2502.

Palmer, T.N. (1995) Predictability of the atmosphere and oceans. In: *Seminar Proceedings of the ECMWF seminar on predictability*, Reading, 4-8 September 1995, Vol. I, pp. 83-141.

Roads, J.O. (1987) Predictability in the extended range. J. Atmos. Sci. 44, 3495–3527.

Roebber, P.J. (1995) Climate variability in a low-order coupled atmosphere-ocean model. Tellus 47A, 473-494.

Roebber, P.J., Bruening, S.L., Schultz, D.M. and Cortinas Jr., J.V. (2003) Improving snowfall forecasting by diagnosing snow density. Wea. Forecasting 18, 264-287.

Tribbia, J.J., and Baumhefner, D.P. (2004) Scale interactions and atmospheric predictability: An updated perspective. Mon. Wea. Rev. 132, 703-713.

Tsonis, A.A. (1992) *Chaos: From theory to applications*. Plenum, New York.

Whitaker, J.S., and Loughe, A.F. (1998) The relationship between ensemble spread and ensemble mean skill. Mon. Wea. Rev. 126, 3292-3302.

29 The Role of El Niño—Southern Oscillation in Regulating its Background State

De-Zheng Sun

CU/CIRES/Climate Diagnostics Center & NOAA/ESRL/Physical Science Division
Boulder, Colorado
Dezheng.Sun@Noaa.Gov

Abstract. A nonlinear aspect of the El Niño—Southern Oscillation (ENSO)—its regulatory effect on the background state (the climatological state)—is described. In particular, it is shown that ENSO acts as a basin-scale heat "mixer" that prevents any significant increase from occurring in the time-mean difference between the warm-pool SST (Tw) and the temperature of the thermocline water (Tc). When this temperature contrast is forced to increase, the amplitude of ENSO increases—El Niño becomes warmer and La Niña becomes colder. A stronger La Niña event results in more heat transported to the subsurface of the western Pacific. A stronger El Niño event then warms the eastern Pacific and cools the western Pacific. The effect of a stronger La Niña event does not cancel the effect from a stronger El Niño event. The long-term mean effect of ENSO—the recurrent occurrence of El Niño and La Niña events—is to mix heat downward across the equatorial Pacific and prevent the time-mean difference between Tw and Tc from exceeding a critical value.

The results have implications for several climatic issues and these implications are discussed. In particular, it is noted that our existing paradigm to understand the response of ENSO to global warming needs to be modified. It is emphasized that it is the tendency in the stability forced by an increase in the greenhouse effect, not the actual changes in the time-mean climate, that ENSO responds to. Changes in the latter—changes in the mean climate—are a residual between the effect of the changes in the radiative forcing and the effect of the changes in the ENSO behavior.

1 Introduction

Among our many questions about the behavior of the Earth's climate system, one stands out prominently. The question is whether natural variability in the state of the climate system—the temporal instability in the climate system—plays a role in regulating the long-term stability of the climate system. (An adjunct question is whether natural variability plays a role in determining the sensitivity and stability of the climate system to anthropogenic forcing). In this context, the role of El Niño—Southern Oscillation (ENSO) in determining the climatological state of the coupled ocean-atmosphere system stands as a fundamental issue in the sciences of global climate change and by extension a fundamental issue in geosciences.

Although the last 20 years has witnessed heightened research activity in understanding the nature of ENSO (Neelin et al 1998, Wang et al. 2004), the nonlinear aspects of ENSO in general, its interaction with the mean background flow in particular, have just began to be revealed. This chapter highlights some progress we have made on this front of research with a focus on the role of ENSO in regulating the time-mean difference between the sea surface temperature of the warm-pool (Tw) and the temperature of the equatorial undercurrent (Tc). Because the water feeding the equatorial undercurrent ultimately comes from the high latitudes, and because the poleward heat transport depends on the vertical mixing of heat in the tropical oceans, we point to a possible role of ENSO in determining the equator-to-pole temperature differences and thereby the mean surface temperature of the planet Earth.

The chapter is organized as follows. We first describe some phenomenological aspects of ENSO and highlight the fact that ENSO corresponds to a basin-wide "sloshing" of the water in the equatorial upper Pacific (Section 2). In section 3, we briefly review the linear theories that have developed to explain the occurrence of the ENSO phenomenon. We will point out there that the linear theories, while quite successful in explaining ENSO as an anomaly about a prescribed background state, left unaddressed the question of whether ENSO in turn plays a role in determining the background state. We then proceed to discuss the background studies that have lead to the present hypothesis about the role of ENSO in regulating its background state, and present the results from coupled experiments that have been explicitly designed to test the hypothesis (Section 4). Finally, we discuss the implications of our results (section 5).

2 Some Observed Hallmarks of ENSO

Even though the incoming solar radiation at the outer boundary of the climate system is zonally symmetric over the equatorial Pacific, the equatorial western Pacific is considerably warmer than the equatorial eastern Pacific (Fig. 29.1a). The warm water in the equatorial western Pacific—the warm water enclosed by the 28^{o}C SST contour in particular—is often referred to as the western Pacific warm-pool. The western Pacific warm-pool has the maximum sea surface temperature of the Earth's open oceans. The western Pacific warm-pool is a fairly deep structure and therefore contains a huge amount of heat (Fig. 29.1b). The warm-pool region acts as a major "furnace" of the climate system because tropical deep convection is concentrated there. The resulting latent release over the warm-pool drives the planetary-scale circulations in the atmosphere which in turn drive the circulations in the oceans (Fig. 29.1c). The cold-SST in the equatorial eastern Pacific results from a wind-driven shoring of the thermocline in that region and the associated cold water upwelled to the surface to feed the surface divergence. The "tongue-like" shape in the SST distribution in the eastern equatorial Pacific has led to the term "equatorial Pacific cold-tongue" (Wyrtki 1981).

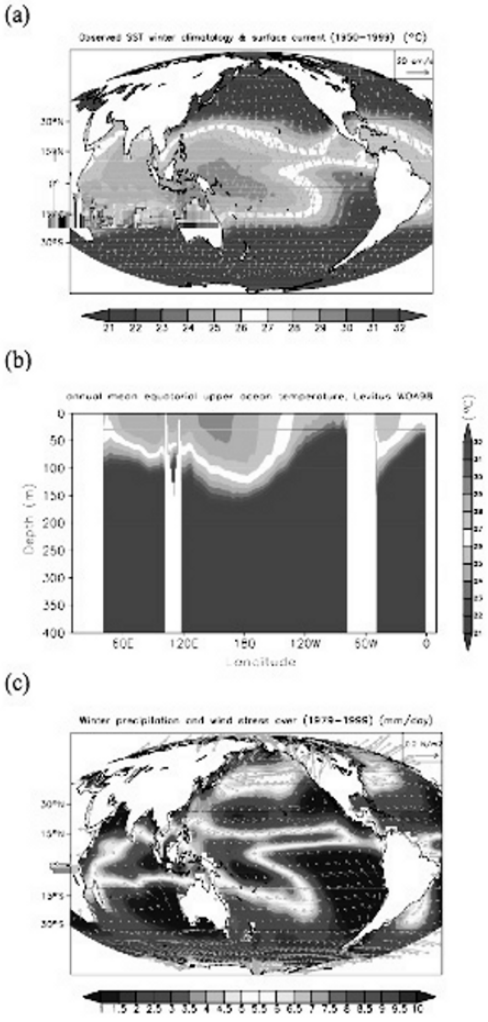


Fig. 29.1. (a) Climatological global SST distribution. The arrows in green show the corresponding surface currents. (b) The corresponding equatorial upper ocean temperature distribution (5°S-5°N). (c) The corresponding distribution of precipitation and surface wind stress. Shown are winter conditions (Nov.-Dec.). The SST and currents data are from the global ocean analysis (Carton et al. 2000). The surface winds data are from the NCEP reanalysis (Kalnay et al. 2000). The precipitation data are from Xie and Arkin (1996).

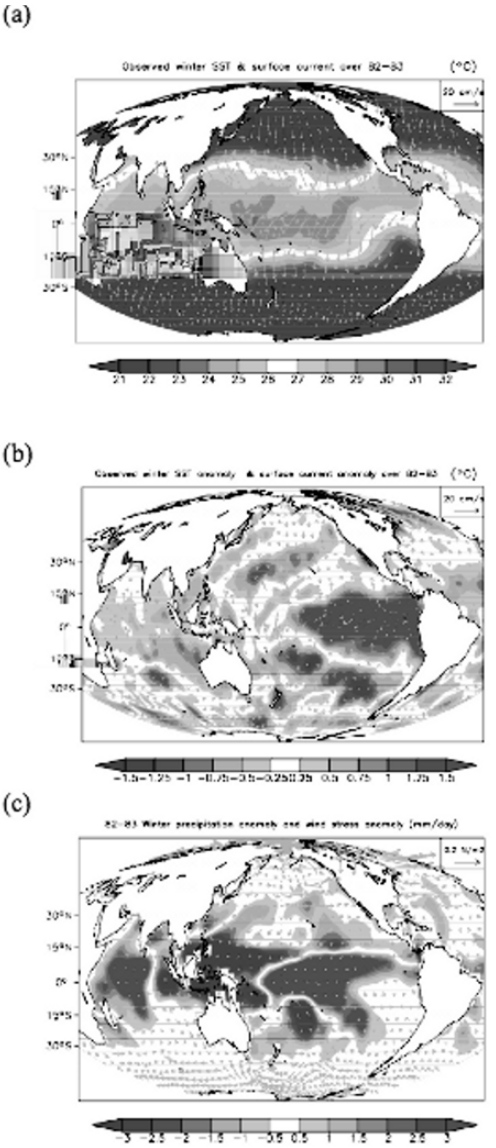


Fig. 29.2. (a) Global SST distribution for the 1982-83 winter. The arrows in green show the corresponding distribution of surface currents. (b) The SST anomaly for the 1982-83 winter. The arrows indicate the corresponding surface current anomaly. (c) The precipitation and surface wind anomalies (arrows) for the 1982-83 winter.

The warm-pool—cold-tongue configuration is not steady. The relative size of the warm-pool and the cold-tongue varies on a range of time-scales. The most profound change occurs on the inter-annual time-scales. Every 2—7 years, the western Pacific warm-pool extends to the east and causes anomalous warming in the central and eastern equatorial Pacific (Fig. 29.2ab). This eastward displacement also displaces the deep convection eastward and is associated with a weakening of the normally eastward trade winds (Fig. 29.2c). This eastward displacement is called El Niño. An El Niño event is often followed by a westward extension of the cold-tongue. The resulting anomalous decrease in the eastern and central equatorial Pacific SST is called La Niña. Accompanying with this alternating extension and retreat of the warm-pool or cold-tongue are the alternating strengthening and weakening of the trade winds. This quasi-periodic eastward extension of the western Pacific warm-pool and the associated atmospheric changes are called El Niño—Southern Oscillation (ENSO). As trade winds upon the equatorial Pacific strengthens and weakens during the ENSO cycle, the upper ocean water is moved back and forth zonally as well as meridionally (Wyrtki 1985, Neelin et al 1999)—ENSO effectively "sloshes" the water in the equatorial upper ocean, particularly in the zonal direction.

3 Explaining ENSO—The Linear Perspective

A milestone in our understanding of ENSO is the numerical model of Zebiak and Cane (1988). This model underscores two important elements for the recurrent occurrence of El Niño events which result in repeatedly "sloshing" of water in the equatorial upper ocean. One is the Bjerknes feedback—the positive feedback loop among anomalous SST, anomalous surface winds, and anomalous upwelling. This positive feedback loop enables an initial surface warming/cooling in the central and eastern Pacific to grow (Neelin et al. 1998).

The model of Zebiak and Cane (1988) is an anomaly model for ENSO, however. The model does not have a heat budget for the subsurface ocean, and consequently had to resort to something that is empirical to wrap up the instability. This empirical fix is a prescribed relationship between the temperature of the upwelling water and the depth of the equatorial thermocline. Nonetheless, the model results suggested another element that is later proven quite relevant to the real system—the delayed negative feedback from the ocean dynamics (Schop and Suarez 1988, Battisti 1988). The nature of the delayed negative feedback is further clarified in the discharge and recharge oscillator theory of Jin (1996) which explicitly links the negative feedback and the associated ocean dynamics to the heat transport process. Jin's theory remains a theory for an anomaly and requires the prescription a background state. The background state is usually chosen as the observed climatological state. The question whether ENSO in return affects the climatological state is left unaddressed.

Another school of thought on the origin of El Niño events is the stochastic theory of ENSO (Penland and Sardeshmukh 1995). Based on linear inverse modeling of ENSO, these authors argue that ENSO can be explained by a linear system forced by noise. They further interpret the noise as a way to represent the unresolved nonlinearity or the weather events or the combination of the two. Again, this view requires

a prescription of a fixed background state, leaving the question whether ENSO in turn plays a role in determining the background state unaddressed.

4 The Effects of ENSO on Its Background State

4.1 Some background material

The theoretical study of Sun and Liu (1996) first suggested the importance of the basin-scale coupling between the atmosphere and ocean in regulating the tropical maximum SST and by extension the importance of the basin-scale coupling in regulating the time-mean tropical Pacific climate. Since ENSO results from the same basin-scale coupling between the atmosphere and ocean, the study of Sun and Liu (1996) heightened the interest in the role of ENSO in the long-term heat balance of the tropical Pacific. Subsequent analysis of the observed heat balance in the tropical Pacific reveals that ENSO is fundamentally involved in the heat transport in the tropical Pacific (Sun and Trenberth 1997, Sun 2000). Sun and Trenberth (1997) noted that during the 1986-87 El Niño event, there is not only marked increase in cloud reflection of solar radiation, but also marked increase in the pole-ward heat transport in the atmosphere and ocean. The extended study by Sun (2000) and Sun (2003) further show that transporting heat poleward away from the equatorial Pacific episodically through El Niño events is actually the norm of the way by which equatorial Pacific pushes pole-ward the heat absorbed over the equatorial Pacific Pacific. The numerical experiments by Sun (2003) and Sun et al. (2004) show that when the coupled tropical ocean-atmosphere is "demanded" to transport more heat from the equatorial Pacific to the higher latitudes through an increase in the surface heating over the tropical ocean or through an increase in the surface cooling over the subtropical ocean, ENSO becomes more energetic. This result confirms the intuitive notion that has been developed from the observational analysis that ENSO corresponds to a heat transport mechanism. The numerical experiments also help to clarify the role of La Niña in the heat transport—it helps to pump heat downward. Not like in the tropical atmosphere where an ascending parcel can become buoyant once it is lifted above the boundary layer, and can therefore continue the ascent in a self-propelled manner, warm surface water in the tropical Pacific has to be pumped downward through strong winds which occur during the La Niña phase. The notion that ENSO acts as heat transfer mechanism is therefore raised. Sun (2004) use the word "heat-pump" to describe this emerging picture about ENSO.

4.2 The focus of this chapter

If ENSO corresponds to a mechanism of downward and poleward heat transport, there is a good chance that ENSO may play a role in stabilizing the time-mean state of the coupled tropical Pacific ocean-atmosphere system. The reasoning here is similar to the reasoning that has led to the hypothesis that the neutrality of the time-mean thermal structure of the tropical troposphere to moist convection is due to moist

convection (Xu and Emanuel 1989). As radiative heating of the surface and cooling of the interior of the atmosphere destabilizes the atmosphere, moist convection takes place to deliver heat from the surface to the free troposphere. Consequently, the mean lapse rate is strongly regulated. Just as there is evidence indicating that the mean lapse rate of the tropical troposphere is strongly regulated, there is also evidence indicating that the stability of the long-term mean state of the coupled tropical ocean-atmosphere system is regulated (Penland and Sardeshmukh 1995, Penland et al. 2000).

A key parameter that determines the stability of the coupled tropical ocean-atmosphere system is the difference between the warm-pool SST (Tw) and the temperature of the equatorial undercurrent (Tc) (Fig. 29.3). This is because a larger difference between Tw and Tc would result in a stronger zonal SST contrast and thereby stronger zonal winds and stronger zonal tilt of the equatorial thermocline (Jin 1996, Sun 1997).

Increased tropical heating tends to raise Tw. Extratropical cooling tends to cool Tc as the water feeding the equatorial undercurrent comes from the extratropics (McCreary and Lu 1994, Pedlosky 1987). Our curiosity is whether the presence of ENSO reduces the sensitivity of Tw-Tc to either an increase in the tropical heating or an increase in the subtropical/extratropical cooling.

Our curiosity about the role of ENSO in determining the equatorial upper ocean thermal stratification is also driven by a hallmark of the ENSO phenomenon, namely the "sloshing" of water in the equatorial upper Pacific—water being moved back and forth in both the zonal and meridional direction during an ENSO cycle. Given this correspondence, it is of interest even from a purely phenomenological perspective to investigate the effect of the recurrent occurrence of ENSO on the time-mean upper ocean temperature structure which is characterized by the time mean difference between Tw and Tc.

In light of the results from the observations, we have the following hypothesis concerning the effects of ENSO on its background state: The recurrent occurrence of

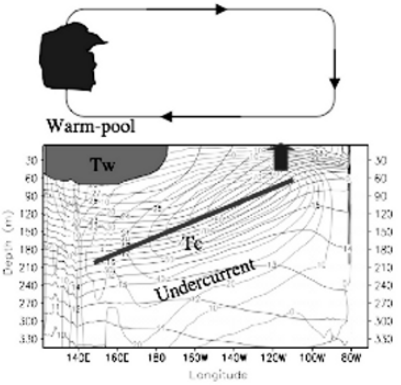


Fig. 29.3. A schematic showing the role of Tw-Tc in the coupled tropical ocean-atmosphere system.

El Niño and La Niña events—ENSO—may be a mechanism that regulates the time-mean thermal stratification of the equatorial upper ocean, and more generally the stability of the long-term mean state of the coupled tropical Pacific ocean-atmosphere system.

4.3 The methodology

Our method to test this hypothesis is through conducting numerical experiments in pairs with a coupled model. In one experiment, the ENSO is turned off. We turn off ENSO in the model by setting the coupling coefficients between the SST and surface winds in the equatorial region to zero. In the other experiment, ENSO is kept on. We then contrast the differences in the response in the time-mean climate, the difference between Tw and Tc in particular as it determines the stability of the time-mean climate.

The perturbation takes the form of either increased radiative heating over the tropics or alternatively increased radiative cooling over the subtropics. For reasons mentioned in the introduction, changes like these tend to increase the difference between Tw and Tc and thereby tends to destabilize the coupled climate system.

The model is the one used in Sun (2003) and Sun et al. (2004). The model has the NCAR Pacific basin model (Gent and Cane 1989) as its ocean component. Thus the model calculates the upper ocean temperatures based on the first principles. This feature is in contrast with the intermediate ocean models used in the pioneering studies of ENSO (Neelin et al. 1998). (Intermediate ocean models do not have a heat budget for the subsurface ocean). The parameterization of the heat flux and zonal wind stress in the present model, however, are in line with those used in previous theoretical studies of ENSO and tropical climatology (Neelin et al. 1998). Specifically, we assume that the surface heat flux is proportional to the difference between the radiative convective equilibrium sea surface temperature (SSTp) and the actual SST. This parameterization is supported by observations (Sun and Trenberth 1998) and aids the changes of radiative heating in the perturbation experiments. The model simulates major characteristics of ENSO including the frequency of occurrence and the subsurface temperature evolution over a life cycle (Sun 2003).

4.4 Tropical heating experiments

Table 29.1 shows the response in Tw, Tc, and Tw-Tc in three pairs of the perturbation experiments. The three pairs are presented here and they differ in the regions where an enhanced tropical heating is applied. In all these pairs, the response in Tw-Tc in the presence of ENSO is much reduced than in the absence of ENSO. Take pair II as an example, change in Tw-Tc in the absence of ENSO is about 1.3 ^{o}C. With ENSO this change is reduced to 0.1 ^{o}C, following a reduction in the increase in Tw and an

Table 29.1. Response of Tw, Tc, and Tw-Tc to an enhanced tropical heating with and without ENSO. The definitions of Tw and Tc are the same as in Sun et al. (2004). The three pairs presented here differ in the meridional extent of the regions where an enhanced tropical heating is applied. Anomalous heating is confined to 5°S-5°N for Pair I, 10°S-10°N for Pair II, and 15°S-15°N for Pair III. In all three cases, the increase in the radiative convective equilibrium SST (SSTp) peaks at the equator with a value of 2°C, and then decrease with latitude following a cosine profile to zero at the specified latitudes (i.e., 5°, 10°, and 15° respectively for the three cases). The last 23 years of a 27-year-long run are used in the calculation for the case with ENSO. For the case without ENSO, the last 3 years of data of a 27-year-long run are used in the calculation because there is little interannual variability in this case.

perturbation type	experiment type	change of Tw (°C)	change of Tc (°C)	change of Tw-Tc (°C)
Pair I (5°S-5°N)	No ENSO	1.03	0.0050	1.02
	With ENSO	0.81	0.76	0.053
Pair II (10°S-10°N)	No ENSO	1.38	0.036	1.34
	With ENSO	0.97	0.83	0.14
Pair III (15°S-15°N)	No ENSO	0.95	0.24	0.71
	With ENSO	0.55	0.83	-0.085

increase in the response of Tc. Apparently, the presence of ENSO reduces the surface warming and increase the response in the subsurface.

Figure 29.4a and Fig. 29.4b provide a more detailed look of the response in the time-mean temperature in the equatorial upper ocean. These two figures show respectively the equatorial temperature differences between the control run and the perturbed run for the case with ENSO and for the case without ENSO. Without ENSO, the warming of the tropical ocean is essentially confined in the surface layer (Fig. 29.4a). With ENSO, the response extends to the thermocline (Fig. 29.4b). The thermocline water is also warmer while the response at the surface level is reduced, particularly in the western Pacific. The presence of ENSO also appears to warm slightly the surface ocean of the central Pacific (180°-240°E) and cools the surface ocean near the eastern boundary.

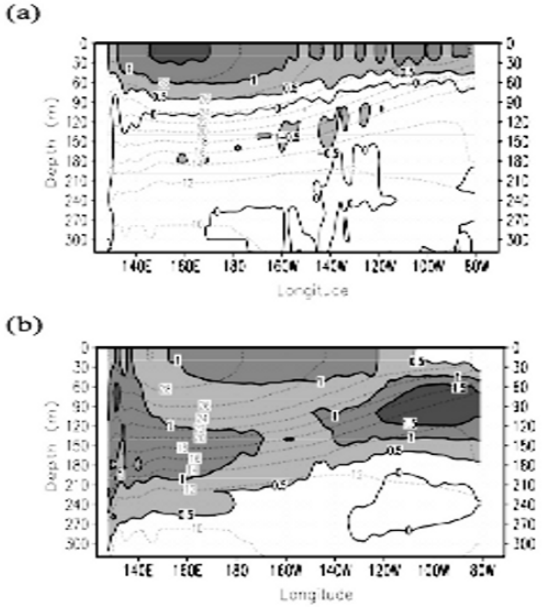


Fig. 29.4. Time-mean equatorial upper ocean temperature response to an enhanced tropical heating for the case without ENSO (a) and the case with ENSO (b). Shown are the results from experiments of Pair II listed in Table 29.1. Data used for the calculations here are the same as used for obtaining changes in Tw and Tc in Table 29.1. The thin dashed contours indicate the mean isentropes of the control run.

To understand these differences, we note that the ENSO is stronger in response to enhanced tropical heating. Fig. 29.5 contrasts the Niño3 SST time-series from the control run with that from the perturbed run. The figure shows that in response to the increase in the tropical heating, El Niño events become warmer while the La Niña events become colder. The variance of the Niño3 SST in the perturbed run is more than four times of the variance of the Niño3 SST in the control run (the standard deviation changes from 0.55 ^{o}C to 1.3 ^{o}C). The changes in the stand deviation of Niño3 SST in the other two pairs of experiments are respectively from 0.55 ^{o}C to 1.35 ^{o}C. Clearly, ENSO in this model is sensitive to tropical heating.

Stronger La Niña enables more heat to be transported downward to the subsurface ocean, the western Part of the ocean in particular in the equatorial region. The top panel on Fig. 29.6 shows the upper ocean temperature changes during La Niña events in response to the enhanced tropical heating. The thermocline is deeper in the western Pacific consistent with the stronger zonal winds. The bottom panel on Fig. 29.6 shows the temperature differences in the warm-phase. Stronger El Niño events then warm the upper ocean in the central and eastern Pacific. Averaged over the cold and warm phases, the heat is "mixed" downward across the basin. Note the asymmetry between changes in the upper ocean temperature during the La Niña phase (Fig. 29.6a) and the El Niño phase (Fig. 29.6b).

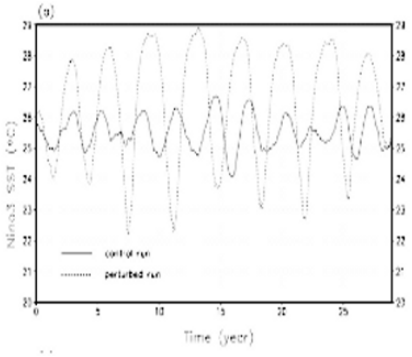


Fig. 29.5. Response of ENSO in the coupled model to an increase in the tropical heating. Shown are time series of Niño3 SST from a control run (solid line) and a perturbed run (dashed line).

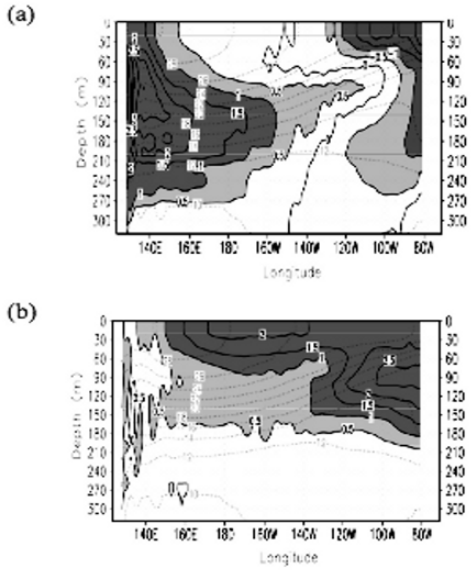


Fig. 29.6. Differences in the equatorial upper ocean temperature between the perturbed run and the control run during the La Niña phase (a) and during the El Niño phase (b). The results are from pair II (see legends of Table 29.1). The 6 cycles of the last 23-years of a 27-year-long run are used in this calculation.

4.5 Extratropical cooling experiments

An important finding of the last decade or so in physical oceanography is that the equatorial thermocline water originates from the extratropical surface ocean. Surface water in the subtropical/extratropical ocean is pumped downward and equatorial ward and eventually feed the equatorial undercurrent and the equatorial upwelling

water (McCreary and Lu 1994, Pedlosky 1987). This tropical extratropical/subtropical connection has been referred to as the "ocean tunnel" and is schematically shown in Fig. 29.7. If you cool the extratropical/subtropical surface ocean, you can cool the equatorial thermocline water and thereby destabilize the coupled system over the equatorial Pacific. But again, we will see that the response in the temperature of the equatorial undercurrent (Tc) to a cooling imposed over the subtropical/extratropical ocean depends critically on whether the system has ENSO or not.

Table 29.2 shows the response in Tw, Tc and Tw-Tc in three pairs of the perturbation experiments. The three pairs are presented here and they differ in the regions where an enhanced subtropical/extratropical cooling is applied. In all these pairs, the response in Tw-Tc in the presence of ENSO is much reduced than in the absence of ENSO. In fact, the value of Tw-Tc in the perturbed cases is even slightly smaller than in the corresponding control case. Take pair II as an example, change in Tw-Tc in the absence of ENSO is about 0.53 ^{o}C. With ENSO this change is -0.22 ^{o}C, following an increase in the cooling Tw and a reduction in the cooling to Tc. Apparently, the presence of ENSO reduces the subsurface cooling and increases the cooling the surface ocean in the western Pacific.

Figure 29.8a and Fig. 29.8b provide a more detailed look of the response in the time-mean temperature in the equatorial upper ocean. These two figures show respectively the equatorial temperature differences between the control run and the perturbed run for the case with ENSO and for the case without ENSO. Without ENSO, the equatorial thermocline water is cooled by about 1 ^{o}C by the imposed cooling over the extratropical surface ocean. With the presence of ENSO, the cooling to the equatorial thermocline water is reduced to about 0.5 ^{o}C.

To understand these differences, we note again that the ENSO is stronger in response to enhanced extratropical cooling. Figure 29.9 shows the Niño3 SST time-series from the control run and the perturbed run. The El Niño events become warmer while the La Niña events become colder. Just as in the enhanced tropical heating case, stronger La Niña enables more heat to be transported downward to the subsurface

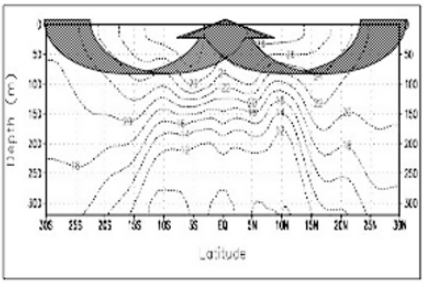


Fig. 29.7. A schematic showing the "ocean tunnel" and how a cooling from the subtropical/extratropical surface may destabilize the equatorial Pacific ocean-atmosphere system.

Table 29.2. Response of Tw,Tc, and Tw-Tc to enhanced subtropical/extratropical cooling with and without ENSO. Three pairs are presented here and they differ in the regions where an enhanced subtropical heating is applied. Pair I corresponds to the case where anomalous subtropical cooling—a reduction in SSTp—starts from 5°. Pair II corresponds to the case where anomalous subtropical cooling starts from 10° Pair III corresponds to the case where anomalous subtropical cooling starts from 15°. In all three cases, the reduction in SSTp increases monotonically with latitude to a fixed value 1°C at 30°S(N). The last 3-year average of a 27-yr-long run are used in the calculation for the case without ENSO. For the three cases with ENSO, the last 23 year of a 36 year long run is used for the case in Pair I, the last 23 years of a 40 year run was used for the case in pair II, and the last 23 years of 56 year long run are used in the calculation for the case in pair III. As the cooling is moved progressively to a higher latitude, there is more delay for the onset of the region with stronger ENSO and consequently there is a need for a longer run to obtain a time series of Niño3 SST that is representative of the regime.

perturbation type	experiment type	change of Tw (°c)	change of Tc (°c)	change of Tw-Tc (°c)
Pair I 30°N(S)-5°N(S)	No ENSO	-0.13	-0.75	0.62
	With ENSO	-0.67	-0.45	-0.22
Pair II 30°N(S)-10°N(S)	No ENSO	-0.10	-0.63	0.53
	With ENSO	-0.49	-0.23	-0.26
Pair III 30°N(S)-15°N(S)	No ENSO	-0.080	-0.67	0.61
	With ENSO	-0.62	-0.25	-0.37

ocean, the western Part of the ocean in particular in the equatorial region. The top panel on Fig. 29.10 shows the upper ocean temperature changes during La Niña events in response to the enhanced extratropical heating. The thermocline water in the western Pacific is actually warmer despite the imposed cooling. The right panel on Fig. 29.10 shows the temperature differences in the warm-phase. Stronger El Niño events then warm the upper ocean in the central and eastern Pacific. Averaged over the cold and warm phases, the cooling to the equatorial thermocline water is significantly reduced.

Figure 29.11 further shows a meridional section of the ocean temperature change at the central Pacific. Again, the left is the case without ENSO, and the right panel is for the case with ENSO. The "tunneling" effect is obvious in the case without ENSO. In the presence of ENSO, the temperature anomalies in the equatorial region are so much reduced that the equatorial region appears to be disconnected from the higher latitudes. This may provide an explanation for why studies using observational data tend to show that decadal temperature anomalies originates from the extratropical surface Pacific largely disappear once they reach near the equatorial region (Schneider et al. 1999). The real world has ENSO which effectively destroys the decadal signals originating from the extratropics though effective mixing in the tropical Pacific. Therefore the apparent disappearance of decadal temperature anomalies upon their reach to the equatorial region in the observations does not suggest an ineffective extratropical influence over the level of ENSO activity.

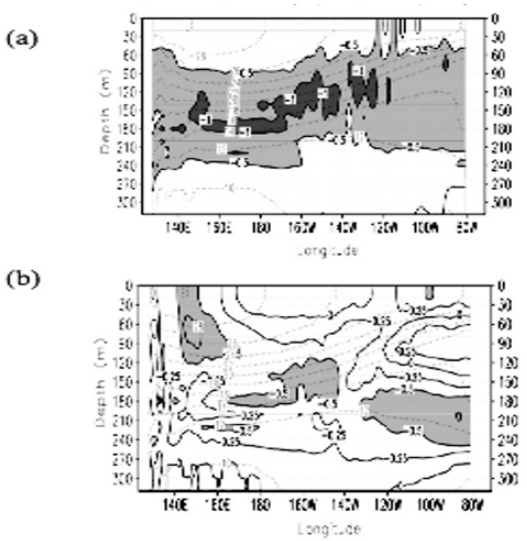


Fig. 29.8. Time-mean equatorial upper ocean temperature response to an enhanced subtropical cooling for the case without ENSO (a) and the case with ENSO (b). Anomalous subtropical cooling—a reduction in SSTp—starts at 10^{o} S(N) for this case and increases monotonically with latitude to a fixed value 1^{o}C at 30^{o}S(N). The last 3 years of data of a 27-yr-long run are used in the calculation for the case without ENSO. For the case with ENSO, the last 23 years of a 40-year long run are used. Not like the almost instantaneous response of ENSO to an increase in the tropical heating, there is a delay for the onset of the regime with stronger ENSO in response to an increase in subtropical cooling. Consequently there is a need for a longer run to obtain a time series of Niño3 SST that is representative of the regime.

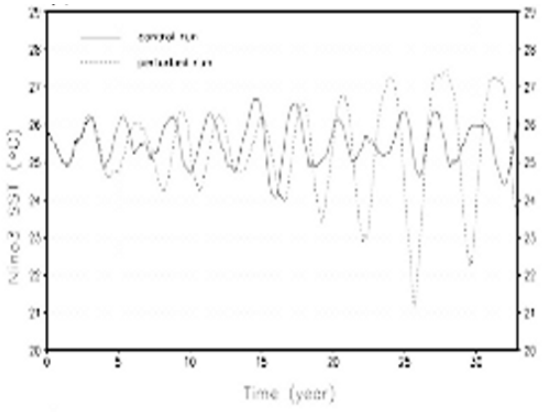


Fig. 29.9. Response of ENSO in the coupled model to an increase in the subtropical/extratropical cooling. Shown are time series of Niño3 SST from a control run (solid line) and a perturbed run (dashed line).

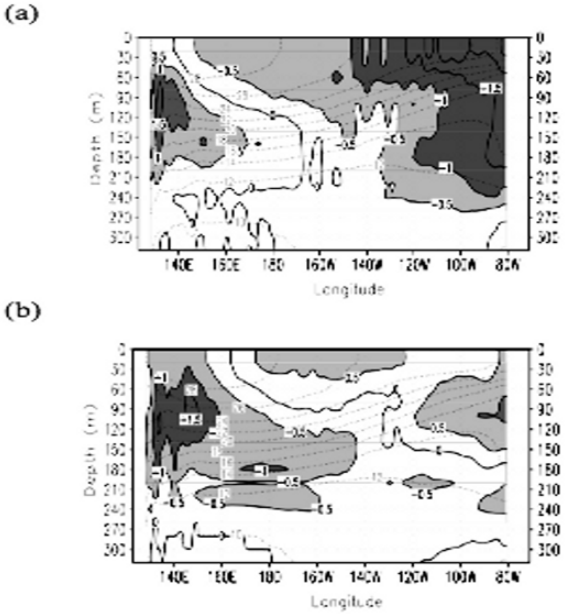


Fig. 29.10. Differences in the equatorial upper ocean temperature between the perturbed run and the control run during the La Niña phase (a) and during the El Niño phase (b). The results are from pair II (see legends of Table 29.2). The 6 cycles of the last 23-years of a 27-year-long run are used in this calculation.

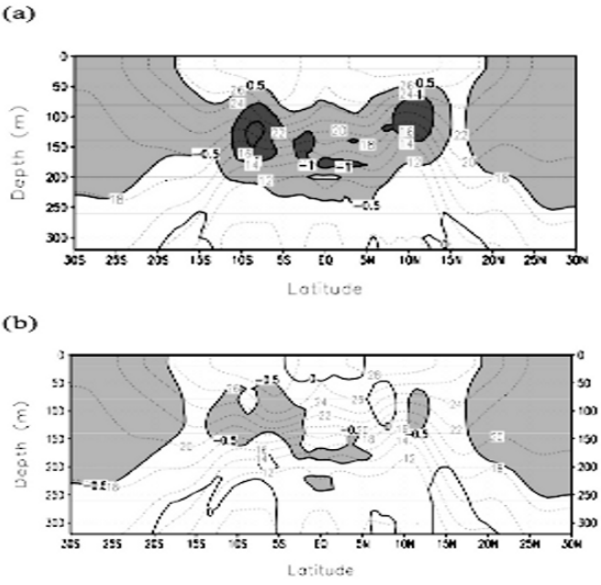


Fig. 29.11. Same as Fig. 8 except for a meridional section for the central Pacific (160 ^{o}E-210^{o}E). (a) Response without ENSO; (b) Response with ENSO

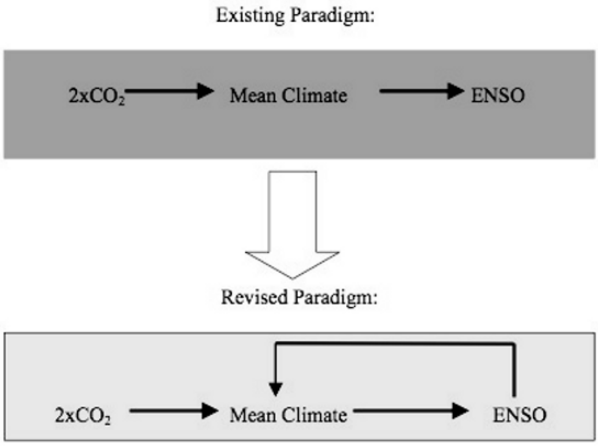


Fig. 29.12. A schematic showing the difference between the revised paradigm and the traditional paradigm.

5 Summary and Discussion

A fundamental question about the Earth's climate system is whether natural variability in the state of the climate system—the temporal instability in the climate system—plays a role in regulating the long-term stability of the climate system. To shed light on this question, we have carried out numerical experiments to investigate the role of ENSO in regulating the stability of the time-mean state of coupled tropical Pacific ocean-atmosphere system. The results suggest that ENSO acts a basin-scale heat mixer that prevents any significant increase from occurring in the time-mean differences between the warm-pool SST (Tw) and the temperature of the thermocline water (Tc). Because the water feeding the equatorial undercurrent ultimately comes from the high latitudes, and because the poleward heat transport depends on the vertical mixing of heat in the tropical oceans, we point to a potential role of ENSO in determining the equator-to-pole temperature differences and thereby the mean surface temperature of the planet Earth.

The finding that ENSO regulates the climatological difference between Tw and Tc has several implications. First, given the efficiency of ENSO in removing heat from the surface ocean to the subsurface ocean, one naturally wonder about the reliability of prediction of global warming made by climate models that do not have good simulations of ENSO. Without adequate ENSO in the models, the heat uptake of the equatorial upper oceans and the poleward heat transport by the oceans may be

underestimated by the models. This prediction may be tested with the output of coupled models.

The results also imply that some biases in the mean climate simulated by the coupled climate models in the tropical Pacific, such as the excessive cold-tongue (Sun et al. 2006), may not just be an cause of errors in the simulated ENSO—it could well be a consequence of the poor simulation of ENSO. The cold-bias in the SST over the central and eastern equatorial Pacific in the coupled models may be indicative of a colder equatorial undercurrent due to an insufficient downward heat mixing by ENSO. This is clearly a conjecture at this stage, but can be tested with the outputs from the coupled models—through analyzing ENSO and the subsurface temperatures of the coupled models. In the same vein, the present finding suggests that the tropical Pacific decadal variability may be more a consequence of the decadal variations in the level of ENSO activity than a cause of them. Therefore, the present results echo the concern raised by Schopf and Burgman (2005). The results also lend support to the suggestion by Rodger et al (2004) and Yeh and Kirtman (2004).

In the same vein, the present finding suggests a need to revise the paradigm that we have used to understand how ENSO may respond to global warming. The traditional paradigm used in understanding the impact of anthropogenic forcing on ENSO is schematically shown in the upper panel of Fig. 29.12. The procedure is that one first finds out the impact of an increase in CO2 on the mean climate—the climatological state, then use the realized changes in the climatological state to deduce the response of ENSO. (For a recent study that epitomizes this procedure, see Fedorov and Philander (2000)). The underlying assumption in this procedure is that changes in ENSO are a passive response to changes in the time-mean climate—ENSO is not involved in determining the changes in the background state. The present finding suggests, however, that ENSO is fundamentally involved in determining at least certain aspects of its background state or the time-mean climate. In particular, the present findings suggest that it is the tendency in the stability forced by increased greenhouse effect, not the actual changes in the time-mean climate, that ENSO responds to. Changes in the latter—changes in the mean climate—are a residual between the effect of the changes in the radiative forcing and the effect of the changes in the ENSO behavior. Indeed, in a recent statistical analysis by Tsonis et al. (2005), it is found that changes in ENSO statistics are correlated with the tendencies in the global mean surface air temperature, but not with the global mean surface air temperature itself. Accumulating evidence suggest that it may be high time for us to modify our paradigm—ENSO needs to be considered at least a significant feedback mechanism in our effort to figure out the consequences of anthropogenic forcing on ENSO and on the climatological state. This revised paradigm is schematically shown in Fig. 29.12.

Finally, we would like to stress the need to conduct additional experiments with more sophisticated models to further test the validity of our findings. The numerical model used for the experiments presented here, is improved compared to the intermediate models, but is still an idealized model. The ENSO in the model is more regular than in the observations, apparently because of the lack of representation of the stochastic aspects of the forcing from the weather events over the warm-pool region. Further studies need to include the stochastic aspects of weather events.

However, as the magnitude of the stochastic forcing from the weather events is likely dependent on the warm-pool SST, we expect that this additional forcing of ENSO is likely to enhance the regulatory effect of ENSO on the mean climate.

Acknowledgement

The research was partially supported by NOAA's CDEP program and partially by NSF's Climate and Large-scale Dynamics program (ATM-9912434, ATM-0331760, and ATM-0553111)

References

Battisti, D.S., 1988: The dynamics and thermodynamics of a warm event in a coupled ocean-atmosphere model, *J. Atmos. Sci,* **45,** 2889-2919**.**

Carton, J.A., G. Chepurin, X. Cao, and B.S. Giese, 2000a: A Simple Ocean Data Assimilation analysis of the global upper ocean 1950-1995, Part 1: methodology, *J. Phys. Oceanogr*., **30**, 294-309.

Fedorov, A.V., and S.G. Philander, 2000: Is El Nino Changing? *Science*, **288**, 1997-2002.

Gent, P.R., and M.A. Cane, 1989: A reduced gravity, primitive equation model of the upper equatorial ocean. *J. Compute. Phys*., **81**, 444-480.

Jin, F.F., 1996: Tropical ocean-atmosphere interaction, the Pacific cold-tongue, and the El Nino-Southern Oscillation. *Science*, **274**, 76-78.

Kalnay, E. and 21 coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437-471.

Neelin, J.D., D.S. Battiti, A.C. Hirst, F.F. Jin, Y. Wakata, T. Yamagata, and S. Zebiak, 1998: ENSO Theory. *J. Geophys. Res*., **103**, 14261-14290.

Penland, C., Flügel, M., and P. Chang, 2000: Identification of Dynamical Regimes in an Intermediate Coupled Ocean-Atmosphere Model. *J. of Climate*, **13**, pp. 2105-2115.

Penland, C, and PD Sardeshmukh, 1995: The Optimal Growth of Tropical Sea Surface Temperature Anomalies. *J. Climate,* **8**, 1999-2024.

Philander, S.G., 1990: El Nino, La Nina, and the Southern Oscillation. Academic Press, New York, 293 pp.

Rodgers, K.B., P. Friederichs, and M. Latif, 2004: Tropical Pacific Decadal Variability and its Relation to Decadal Modulations of ENSO. *J. Climate,* **17**, 3761-3774.

Schopf, P. and R. Burgman, 2005: A Simple Mechanism for ENSO Residuals and Asymmetry. *J. Climate*, Accepted.

Schneider, N., A. J. Miller, M.A. Alexander, C. Deser, 1999: Subduction of Decadal North Pacific Temperature Anomalies: Observations and Dynamics, *J. Phys. Oceanogr.,* **29**, 1056-1070.

Shin, S.-I and Z. Liu, 2000: Response of the equatorial thermocline to extratropical buoyancy forcing. *J. Phys. Oceanogr.*, **30**, 2883-2905.

Sun, D.-Z., 1997: El Niño: a coupled response to radiative heating? *Geophys. Res. Lett.*, **24**, 2031-2034.

Sun, D.-Z., 2000: The heat sources and sinks of the 1986-87 El Niño, *J. Climate*, **13**, 3533-3550.

Sun, D.-Z., 2003: A Possible Effect of An Increase in the Warm-pool SST on the Magnitude of El Niño Warming. *J. Climate*, **16**, 185-205.

Sun, D.-Z. and Z. Liu, 1996 : Dynamic ocean-atmosphere coupling: a thermostat for the tropics. *Science*, **272**, 1148-1150.

Sun, D.-Z. and K.E. Trenberth, 1998: Coordinated heat removal from the equatorial Pacific during the 1986-87 El Niño. *Geophys. Res. Lett.*, **25**, 2659-2662.

Sun, D.-Z., T. Zhang, C. Covey, S. Klein, W.D. Collins, J.J. Hack, J.T. Kiehl, G.A. Meehl, I.M. Held, and M. Suarez, 2005: Radiative and Dynamical Feedbacks Over the Equatorial Cold-tongue: Results from Nine Atmospheric GCMs. *J. Climate* , **19** , 4059-4074.

Sun, D.-Z., T. Zhang, and S.-I. Shin, 2004 : The effect of subtropical cooling on the amplitude of ENSO: a numerical study. *J. Climate,* **17**, 3786-3798.

Tsonis, A.A. ,J.B. Elsner, A.G. Hunt and T.H. Jagger, 2005: Unfolding the relation between global temperature and ENSO. *Geophys. Res. Lett*. doi:10.1029/ 2005GL022875.

Wang, C., S.-P. Xie, and J. A. Carton, 2004: A global survey of ocean-atmosphere interaction and climate variability. In: *Earth's Climate: The Ocean-Atmosphere Interaction.* C. Wang, S.-P. Xie, and J. A. Carton, Eds., AGU Geophysical Monograph.

Wyrtki, K., 1981: An estimate of equatorial upwelling in the Pacific. *J. Phys. Oceanogr.,* **11** (9), 1205-1214.

Wyrtki, K., 1985. Water displacements in the Pacific and the genesis of El Niño cycles, *J. Geophys. Res.-Oceans*, **90**, 7129-7132.

Xie, P., and P.A. Arkin, 1996: Analyses of Global Monthly Precipitation Using Gauge Observations, Satellite Estimates, and Numerical Model Predictions. *J. Climate,* 9, 840-858.

Xu, K.M., and K. Emanuel, 1989: Is the tropical atmosphere conditionally unstable? Mon. Wea. Rev., 117, 1471-1479.

Yeh, S.-W. and B.P. Kirtman, 2004: Tropical decadal variability and ENSO amplitude modulations in a CGCM. *J. Geophys. Res.,* **109**, doi:10.1029/2004JC002442.

30 Nonlinear Dynamics of Natural Hazards

Donald L. Turcotte[1], Sergey G. Abaimov[1], Robert Shcherbakov[1,2], John B. Rundle[1,2]

[1]University of California - Davis, Department of Geology, turcotte@geology.ucdavis.edu, abaimov@geology.ucdavis.edu
[2]University of California - Davis, Center for Computational Science and Engineering, roshch@cse.ucdavis.edu, jbrundle@ucdavis.edu

Abstract. In this paper we consider the nonlinear dynamics of several natural hazards and related models. We will focus our attention on earthquakes, landslides, and forest fires. These are clearly complex phenomena but they exhibit self organization. A consequence of this self organization is scaling laws. We will consider frequency-magnitude statistics and recurrence-time statistics. The frequency-magnitude distributions are power-law and we give a cascade model of cluster coalescence to explain this behavior. The return-time distributions are well approximated by the Weibull distribution. An important characteristic of the Weibull distribution is that it is the only distribution that has a power-law hazard function.

1 Introduction

Bak, Tang, and Wiesenfeld (1988) conceived the concept of self-organized criticality in terms of the sand-pile model. Two other models that exhibit similar behavior are the slider-block model (Carlson and Langer 1989) and the forest-fire model (Drossel and Schwabl 1992). The "avalanches" in each model exhibit power-law frequency-area statistics in a very robust manner with a power-law exponent near unity. An important natural hazard is associated with each of these models: landslides with the sand-pile model, earthquakes with the slider-block model, and forest fires with the forest-fire model. In this paper we will review the frequency-area behavior of these hazards and models and will consider the interval-time statistics of the events.

1.1 Fractals

A near universal feature of many natural hazards is that they satisfy power-law (fractal) frequency-magnitude statistics. The concept of fractals was introduced by Mandelbrot (1967) to explain the scaling properties of the rocky coastline of England. A working definition of a fractal distribution is

$$N = Cr^{-D} \tag{1}$$

where N is the number of objects with a linear dimension greater than r, C is a constant, and D is the fractal dimension. It should be emphasized that a fractal distribution is not a statistical distribution in the usual sense in that an integral of the distribution from $0 < r < \infty$ diverges, the distribution must be terminated for either small or large r or both. Thus fractal distributions are not accepted by statisticians. A fractal distribution implies scale invariance over the range of applicability and fractal distributions have found wide applicability in the geosciences (Korvin 1992; Turcotte 1997).

In this paper we will consider the applicability of fractal statistics to earthquakes, wildfires, and landslides. We will show how models (simulations) produce fractal distributions and provide an explanation for the behavior based on a scale-invariant cascade of cluster coalescence.

1.2 Frequency-magnitude distribution of earthquakes

The classic application of a power-law (fractal) frequency-magnitude distribution for a natural hazard is to earthquakes. For over 50 years, it has been accepted that earthquakes universally obey Gutenberg–Richter scaling; the cumulative number of earthquakes per year in a region with magnitude greater than m, $\dot{N}_{CE}$, is related to m by

$$\log \dot{N}_{CE} = -bm + \log a \tag{2}$$

The constant b is known as the b-value and has a near universal value $b = 0.90 \pm 0.15$. The constant a is a measure of the intensity of the regional seismicity.

When Eq. 2 is expressed in terms of the earthquake rupture area, A_E, instead of earthquake magnitude, this relation becomes a power law (Turcotte 1997)

$$\dot{N}_{CE} \propto A_E^{-b} \tag{3}$$

This is equivalent to the fractal scaling given in Eq. 1 taking $D = 2b \approx 1.8$.

Earthquakes satisfy fractal (power-law) scaling in a very robust manner, despite their complexity. Earthquakes occur in regions of high and low heat flow, with compressional, shear, and tensional stresses, yet they universally satisfy the power-law scaling given in Eq. 3.

As an example of the validity of Gutenberg-Richter scaling we consider seismicity in southern California. The frequency-magnitude distributions of the regional seismicity in southern California on a yearly basis are plotted in Fig. 30.1. For each individual year between 1983 and 2002 the cumulative number of earthquakes N_{GR} with magnitudes greater than m is plotted as a function of m. The period 1983-2003 taken together results in the Gutenberg-Richter power law relation with $b = 1.0$ and $a = 5.4$, shown as the solid straight lines in Fig. 30.1. In Fig. 30.1, there is generally good agreement between each individual year's data and the Gutenberg-Richter relation for the period 1983-2003. The exceptions can be attributed to the aftershock sequences of the 1987 Whittier-Narrows, 1992 Landers, 1994 Northridge, and 1999 Hector Mine earthquakes. With aftershocks removed, the background seismicity in southern California illustrated in Fig. 30.1 is nearly uniform from year to year and is not a function of time. Small earthquakes behave like thermal background noise.

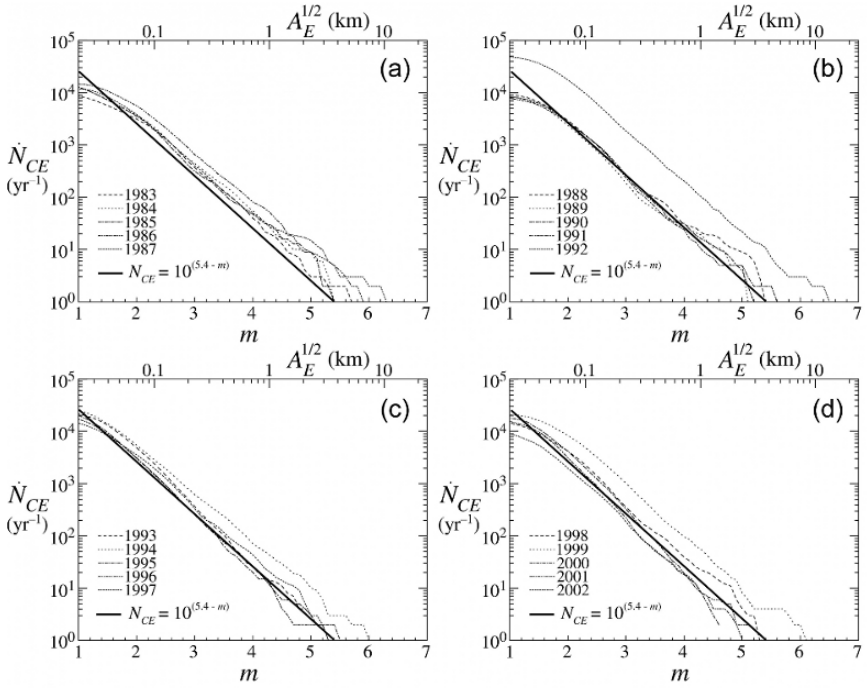


Fig. 30.1. Cumulative number of earthquakes per year, N_{GR}, occurring in southern California with magnitudes greater than m as a function of m (after Rundle, Turcotte, Shcherbakov, Klein, and Sammis (2003)). Twenty individual years are considered. The solid straight line is the Gutenberg–Richter relation (2) with $b = 1.0$ and $a = 5.4$. The larger numbers of earthquakes in 1987, 1992, 1994, and 1999 can be attributed to the aftershocks of the Whittier-Narrows, Landers, Northridge, and Hector Mine earthquakes. If aftershocks are excluded, the background seismicity in southern California is nearly uniform in time.

2 Landslides

Another demonstration of the self organization of natural hazards is the statistical distribution of landslide areas in a landslide event. Landslide events include landslides that are initiated by a trigger such as an earthquake, a large storm, or a rapid snowmelt. A landslide event may be quantified by the frequency-area distribution of the triggered landslides. It has been shown (Malamud, Turcotte, Guzzetti, and Reichenbach 2004) that the frequency-area statistics of substantially complete landslide inventories are well approximated by the three-parameter inverse-gamma distribution.

The probability densities $p(A_L)$ for three landslide inventories are given in Fig. 30.2. A detailed discussion of each inventory is found in Malamud et al. (2004). The three sets of probability densities given in Fig. 30.1 exhibit a characteristic shape, densities increase to a maximum value and then decrease with a power-law tail. Based on the good agreement between these three sets of probability densities, Malamud et al. (2004) proposed a general probability distribution for landslides, a three-parameter inverse-gamma distribution given by

$$p(A_L) = \frac{1}{a\Gamma(\rho)} \left[\frac{a}{A_L - s} \right]^{\rho+1} \exp\left[-\frac{a}{A_L - s} \right] \tag{4}$$

where $\Gamma(\rho)$ is the gamma function of ρ. The inverse-gamma distribution has a power-law decay with exponent $-(\rho + 1)$ for medium and large areas and an exponential rollover for small areas. The maximum likelihood fit of Eq. 4 to the three data sets in Fig. 30.2 yields $\rho = 1.40$, $a = 1.28\times10^{-3}$ km^2, and s = 1.32×10^{-4} km^2, the power-law tail has an exponent $\rho + 1 = 2.40$. Many authors (see Malamud et al. (2004) for a review) have also noted that the frequency-area distributions of large landslides correlate well with a power-law tail. This common behavior is observed despite large differences in landslide types, topography, soil types, and triggering mechanisms.

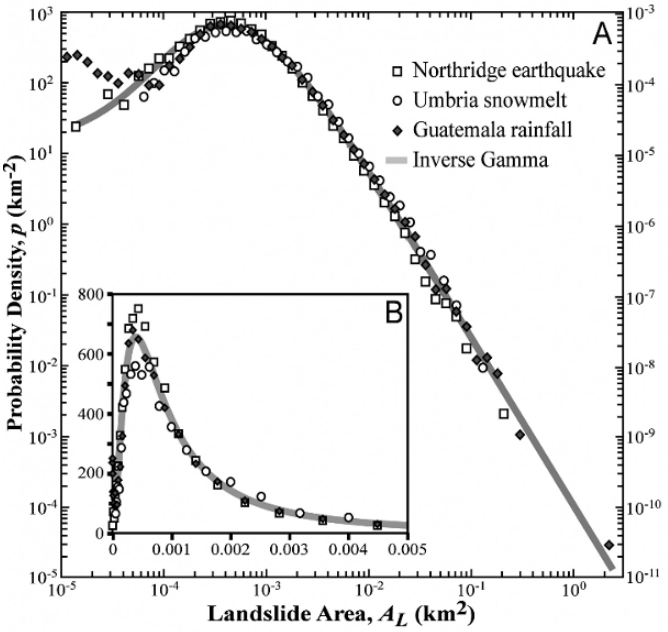


Fig. 30.2. Dependence of landslide probability densities p on landslide area A_L, for three landslide inventories (figure after Malamud et al. (2004)): (i) 11,111 landslides triggered by the 17 January 1994 Northridge earthquake in California (Harp and Jibson 1995); (ii) 4,233 landslides triggered by a rapid snowmelt in the Umbria region of Italy in January 1997 (after Guzzetti, Malamud, Turcotte, and Reichenbach (2002)); (iii) 9,594 landslides triggered by the heavy rainfall from Hurricane Mitch in Guatemala in late October and early November 1998 (Bucknam, Coe, Chavarria, Godt, Tarr, Bradley, Rafferty, Hancock, Dart, and Johnson 2001). Probability densities are given on logarithmic axes in A and linear axes in B. Also included is the general landslide probability distribution. This is the best-fit to the three inventories of the three-parameter inverse-gamma distribution (Eq. 4).

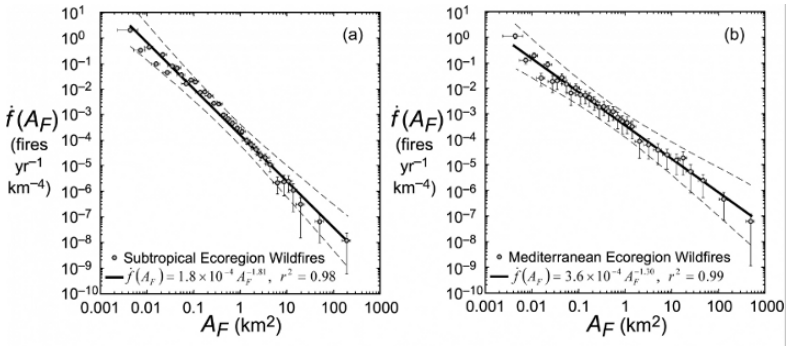


Fig. 30.3. Normalized frequency–area wildfire statistics for (a) Mediterranean and (b) Subtropical ecoregion divisions, for the period 1970–2000 (after Malamud, Millington, and Perry (2005)). Frequency densities $f(A_F)$ are plotted as a function of wildfire area A_F. Also shown is a solid line, the least-squares fit to Eq. 1, with coefficient of determination r^2. Dashed lines represent 95% confidence intervals. Horizontal and vertical error bars represent two standard deviations.

3 Wildfires

Our final example of a natural hazard that follows power-law frequency–area statistics are wildfires. Malamud, Morein, and Turcotte (1998) considered four wildfire data sets from the USA and Australia. In each case the noncumulative frequency-area distribution A_F correlated well with the power-law relationship given in Eq. 3, with b = 1.3-1.5. Malamud et al. (2005) carried out a comprehensive study of the frequency–area statistics of 88,916 wildfires on United States Forest Service lands during the period 1970-2000. For each of eighteen different ecoregions examined they found that the noncumulative number of fires per year plotted as a function of burned fire area A_F correlated well with Eq. 3, with b = 1.30-1.81; two examples are given in Fig. 30.3. In Fig. 30.3a the frequency–area statistics for 16,423 wildfires in the *Subtropical* ecoregion division (within the southeastern part of the USA) are given and in Fig. 30.3b, 475 wildfires in the *Mediterranean* ecoregion division (within California, USA) are given.

A number of other authors (e.g., Minnich (2001); Niklasson and Granstrom (2000); Ricotta, Arianoutsou, Diaz-Delgado, Duguy, Lloret, Maroudi, Mazzoleni, Moreno, Rambal, Vallejo, and Vazquez (2001)) have also found good correlations of the frequency–area distributions of wildfires with the power-law relation. Considering the many complexities of the initiation and propagation of wildfires, it is remarkable that the frequency–area statistics are similar under a wide variety of environments. The proximity of combustible material varies widely. The behavior of a particular wildfire depends strongly on meteorological conditions. Fire-fighting efforts extinguish many fires. Despites these complexities, the power-law frequency–area statistics of actual wildfires seems very robust.

4 Forest-fire Model

The objective of theory, models, and simulations in the earth sciences is to better understand and forecast observed phenomena, specifically natural hazards. There are basically three approaches:

1) Use of deterministic partial differential equations, i.e. the wave equation for the propagation of seismic waves.
2) Statistics, i.e. the use of the normal (Gaussian) distribution for purely random processes.
3) Simulations applied to complex phenomena.

The frequency-magnitude distributions of natural hazards considered in this paper are statistical in nature. Thus solutions of deterministic partial differential equations are not applicable. But the applicable statistics are generally well approximated by fractal (power-law) distributions which find little use in the statistics literature. An example are the simulations that are used to study the behavior of self-organizing complex phenomena. The original simulation was the sand-pile model proposed by Bak et al. (1988). This was the basis for the concept of self-organized criticality.

The forest-fire model (Drossel and Schwabl 1992) was not the first model that was associated with self-organized critical behavior, but it is probably the most illustrative. This model consists of a square grid of sites. At each time step, a model tree is dropped on a randomly chosen site; if the site is unoccupied, the tree is planted. The firing frequency, f, is the inverse number of attempted tree drops on the square grid before a model match is dropped on a randomly chosen site. If $f = 1/100$, there have been 99 attempts to plant trees before a match is dropped at the 100^{th} time step. If the match is dropped on an empty site, nothing happens. If it is dropped on a tree, the tree ignites and a model fire consumes that tree and all adjacent trees. This model is illustrated in Fig. 30.4.

For large time intervals, the number of trees lost in "fires" is approximately equal to the number of trees planted. However, the number of trees on the grid will fluctuate. The frequency-area distribution of "fires" is a statistical measure of the behavior of the system. This model is probabilistic (stochastic) in that the sites are chosen randomly. It is a cellular-automata model in that only nearest-neighbor trees are ignited by a "burning" tree.

Having specified the size of the square grid, N_g, and the firing frequency, f, a simulation is run for N_S time steps and the number of fires N_F with area A_F is determined. The area, A_F, is the number of trees that burn in a fire. Examples of four typical model fires are given in Fig. 30.5. In these examples, the grid size is 128×128, $1/f = 2{,}000$, and fires with $A_F = 5$, 51, 506, and 5,327 trees are illustrated.

The noncumulative frequency-area distributions for model forest fires are given in Fig. 30.6. The number of fires per time step with area A_F,N_F/N_S, is given as a function of A_F. Results are given for a grid size 128×128 and three firing frequencies $1/f = 125$, 500, and 2000. The smaller fires correlate well with the power-law (fractal) relation with exponent between 1.05 and 1.11.

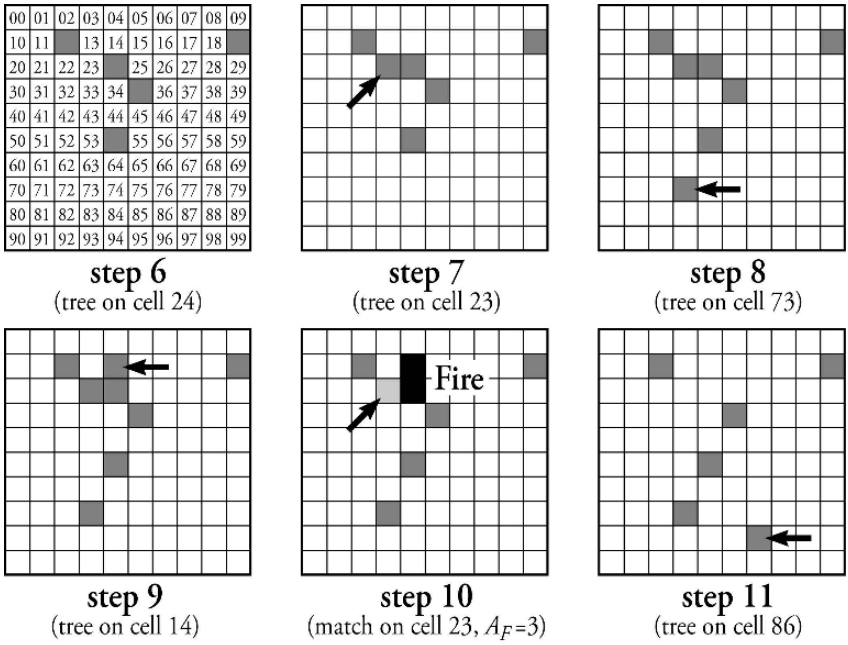


Fig. 30.4. Illustration of the forest-fire model with firing frequency $f = 1/10$. Sites are selected randomly (after Millington, Perry, and Malamud (2006)). In steps 7 to 9 trees are planted on sites 23, 73, and 14. In step 10 a match is dropped on site 23 and trees on sites 14, 23, and 24 burn and are removed from the grid. In step 11 a tree is planted on site 86.

The results for large fires are influenced by the finite-size effect of the grid. If the firing frequency is small, the frequency-area distribution deviates significantly from a straight-line, and there is an upper termination to the power-law distribution. It is seen that large forest fires become dominant for very small firing frequencies. This is easily explained, with a high firing frequency trees burn before large clusters can form. For a small firing frequency clusters form that span the grid before ignition occurs.

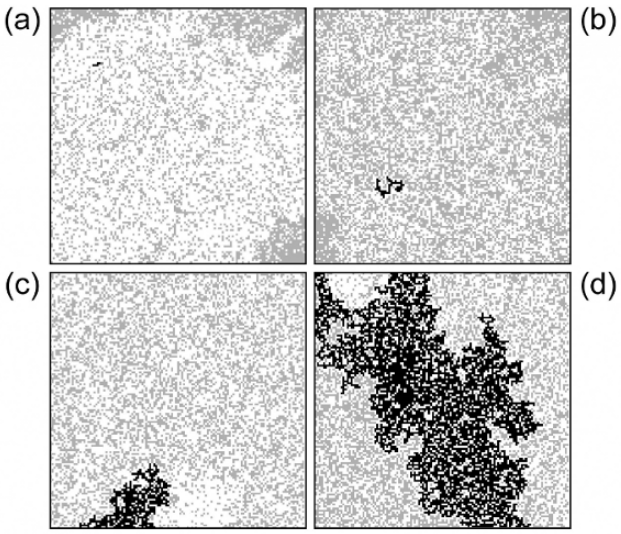


Fig. 30.5. Four examples of typical model forest fires are given for a 128×128 grid with firing frequency $f_S = 1/2{,}000$ (after (Malamud and Turcotte 2000)). The black regions are forest fires. The gray regions are unburned trees. The white regions are unoccupied sites. (a) 5 trees burn, (b) 51 trees burn, (c) 505 trees burn, (d) 5327 trees burn.

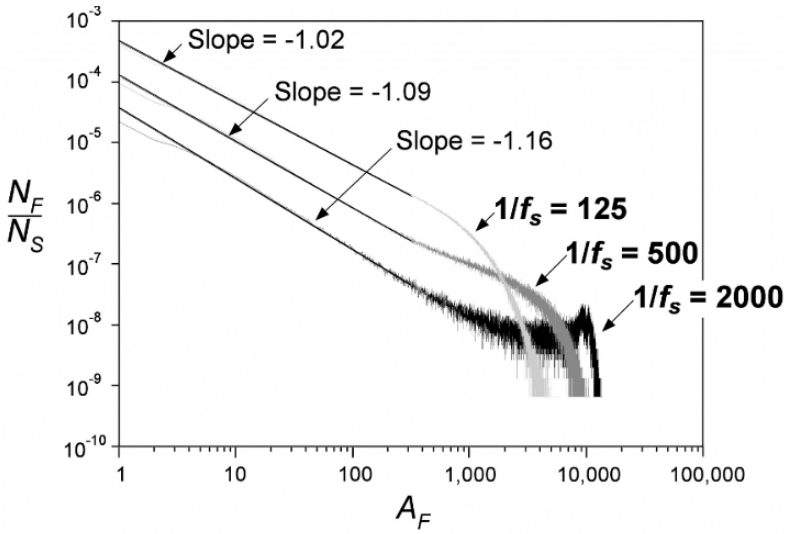


Fig. 30.6. Noncumulative frequency-area distributions of model forest fires (after (Malamud et al. 1998)). The number of fires per time step with area A_F, N_F/N_S, are given as a function of A_F where A_F is the number of trees burned in each fire. Results are given for a 128×128 grid with firing frequency $1/f = 125$, 500, and 2000 and grid sizes 128×128. For each f_S the model was run for $N_S = 1.638\times10^9$ time steps. The small and medium fires correlate well with the power-law relation (4) with $\alpha = 1.02$-1.16. The finite grid-size effect can be seen at the smallest firing frequency.

5 Cascade Model

The scaling behavior of both the forest-fire model and actual forest fires can be understood in terms of a cascade model (Gabrielov, Newman, and Turcotte 1999; Turcotte 1999). This model is formulated in terms of individual trees on a grid and the clusters of trees that are generated. Individual trees are planted on the grid. A new tree can (i) create a new single-tree cluster, (ii) increase the size of an existing cluster by one if planted on a site immediately adjacent to the cluster, or (iii) bridge the gap between two clusters with n_1 and n_2 trees to give a new cluster with n_1+n_2+1 trees. Small clusters coalesce to form larger clusters. Individual trees cascade from smaller to larger clusters. A detailed study of cluster growth in the forest-fire model by Yakovlev, Newman, Turcotte, and Gabrielov (2005) has shown that the last method, cluster coalescence, dominates the cluster growth process. Trees cascade from smaller to larger clusters until they are lost in the fires that destroy the largest clusters, and the cascade is terminated. Fires also destroy smaller clusters, but the largest fires dominate the loss of trees.

The coalescence process is scale invariant and it is easily shown that the frequency-area distribution of clusters is power-law (fractal). It follows that the frequency-area distribution of model fires is also power-law (fractal).

6 Weibull Distribution

Another important aspect of natural hazards is the statistical distribution of interevent times. Specifically, do events such as earthquakes occur randomly in time or are they correlated (aftershocks) or more nearly periodic ("characteristic" earthquakes). Recent work by (Yakovlev, Turcotte, Rundle, and Rundle 2006) indicates that the Weibull distribution may be universally applicable to interevent time statistics in the same way that power-law (fractal) distributions are applicable to frequency-magnitude statistics.

The Weibull distribution is often used for the distribution of life times in engineering applications (Meeker and Escobar 1991; Weibull 1951). Examples include composite materials and fiber bundles. The probability distribution function (pdf) for a Weibull distribution is given by (Patel, Kapadia, and Owen 1976)

$$p(t) = \frac{\beta}{\tau}\left(\frac{t}{\tau}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\tau}\right)^{\beta}\right] \tag{5}$$

where β and τ are fitting parameters. The mean μ and the coefficient of variation C_V of the Weibull distribution are given by

$$\mu = \tau \Gamma\left(1 + \frac{1}{\gamma}\right) \tag{6}$$

$$C_V = \left\{ \frac{\Gamma\left(1 + 2/\gamma\right)}{\left[\Gamma\left(1 + 1/\gamma\right)\right]^2} - 1 \right\}^{\frac{1}{2}} \tag{7}$$

where $\Gamma(x)$ is the gamma function of x. The coefficient of variation C_V is the ratio of the standard deviation σ to the mean μ. The cumulative distribution function (cdf) for the Weibull distribution is given by

$$P(t) = 1 - \exp\left[-\left(\frac{t}{\tau}\right)^{\beta} \right] \tag{8}$$

where $P(t)$ is the fraction of the recurrence times that are shorter than t.

If $\beta = 1$ the Weibull distribution becomes the exponential (Poisson) distribution with $\sigma = \mu$ and $C_V = 1$. In the limit $\beta \to +\infty$ the Weibull distribution becomes a δ-function with $\sigma = C_V = 0$. In the range $0 < \beta < 1$ the Weibull distribution is often referred to as the stretched exponential distribution.

An important property of the Weibull distribution is the power-law behavior of the hazard function

$$h(t_0) = \frac{\mathrm{p}(\mathrm{t}_0)}{1 - P(t_0)} = \frac{\beta}{\tau}\left(\frac{t_0}{\tau}\right)^{\beta - 1} \tag{9}$$

The hazard function $h(t_0)$ is the pdf that an event will occur at a time t_0 after the occurrence of the last event. For the Poisson distribution, $\beta = 1$, the hazard function is constant $h(t_0) = \tau^{-1}$ and there is no memory of the last event. For $\beta > 1$ the hazard rate increases as a power of the time t_0 since the last event. The hazard rate is expected to increase monotonically for a driven system.

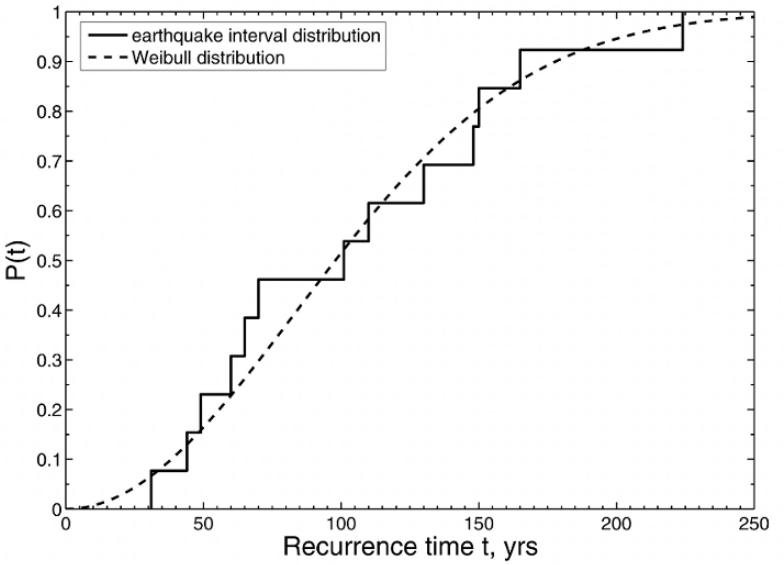


Fig. 30.7. Cumulative distribution of recurrence times of great earthquakes at the Wrightwood site on the southern San Andreas fault. The solid line is the distribution of 13 actual recurrence times. The dashed line is the best-fit Weibull distribution ($\mu = 103.6$ years, $C_V = 0.532$, $\tau = 117.5$ years, and $\beta = 2.01$).

7 Southern San Andreas Fault

Displacements on the southern section of the San Andreas fault occur in great earthquakes. Using paleoseismic studies Biasi, Weldon, Fumal, and Seitz (2002) have quantified the recurrence times at the Wrightwood site on the San Andreas fault and found that the interval times between great earthquakes are 110, 65, 31, 60, 70, 165, 101, 148, 224, 49, 150, 130, and 44 years. The mean and coefficient of variation of these recurrence times are $\mu = 103.6$ years and $C_V = 0.532$. The cumulative distribution of recurrence times $P(t)$ is given in Fig. 30.7. Also included in Fig. 30.7 is the best log likelihood (-69.55) fit of the Weibull distribution with $\tau = 117.5 \pm 17.1$ years and $\beta = 2.01 \pm 0.44$ (the error bars are 95% confidence limits).

8 Virtual California Simulations

Ideally the recurrence statistics of characteristic earthquakes would be used to determine the applicable statistical distribution. Although the fit of the Weibull distribution given above is quite good, the number of events is not sufficient to establish definitively its validity (Savage 1994).

In order to overcome this lack of data, several numerical simulations of earthquake statistics have been carried out. We give results using the Virtual California simulation (Rundle, Rundle, Donnellan, and Fox 2004; Rundle, Rundle, Donnellan, Turcotte, Shcherbakov, Li, Malamud, Grant, Fox, McLeod, Yakovlev, Parker, Klein, and Tiampo 2005; Yakovlev et al. 2006). This model is a geometrically realistic numerical simulation of earthquakes occurring on the San Andreas fault system. It includes the major strike-slip faults in California and is composed of 650 fault segments, each with a width of 10 km and a depth of 15 km. The fault segments interact with each other elastically utilizing dislocation theory. Virtual California is a backslip model, the accumulation of a slip deficit on each segment is prescribed using available data. The mean recurrence times of earthquakes on each segment are also prescribed using available data to give friction law parameters. The statistical distribution of recurrence times on the northern San Andreas fault (site of the 1906 San Francisco earthquake) was obtained from a 1,000,000 year simulation. The mean recurrence time for 4606 simulated earthquakes with $M > 7.5$ on this section is $\mu = 217$ years and the coefficient of variation is $C_V = 0.528$. The cumulative distribution of recurrence times $P(t)$ is given in Fig. 30.8. Also included in Fig. 30.8 is the best fit of the Weibull distribution with $\tau = 245$ years and $\beta = 1.97$. Details of the quality of fit have been given by Yakovlev et al. (2006).

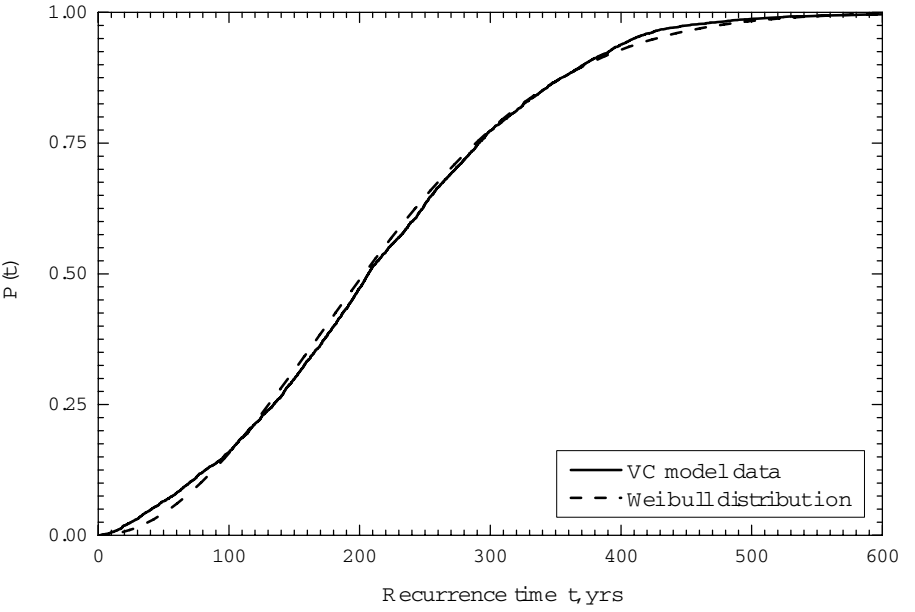


Fig. 30.8. Cumulative distribution of recurrence times for 4606 simulated earthquakes with $M > 7.5$ on the northern San Andreas fault. The solid line is the distribution of simulated recurrence times. The dashed line is the best-fit Weibull distribution ($\mu = 217$ years, $\sigma = 115$ years, $C_V = 0.528$, $\tau = 245$ years, and $\beta = 1.97$).

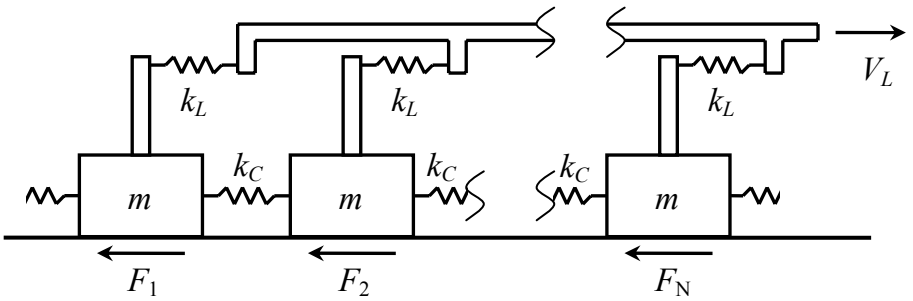


Fig. 30.9. Illustration of the one-dimensional slider-block model. A linear array of N blocks of mass m are pulled along a surface by a constant velocity V loader plate. The loader plate is connected to each block with a loader spring with spring constant k_L and adjacent blocks are connected by springs with spring constant k_C. The frictional resisting forces are $F_1, F_2, \ldots, F_N$.

9 Slider-Block Model

Burridge and Knopoff (1967) proposed the multiple slider-block model as a simple model for earthquake occurrence. Carlson and Langer (1989) showed that the multiple slider-block model exhibits self-organized criticality. The frequency-size distribution of slip events is fractal as is the case for earthquakes (Turcotte 1997). We will now consider the interval statistics of slip events. We will obtain both the Poisson behavior of regional seismicity and the Weibull behavior of characteristic earthquakes. We utilize a variation of the linear slider-block model which Carlson and Langer (1989) used. We consider a linear chain of 100 slider blocks of mass m pulled over a surface at a constant velocity V_L by a loader plate as illustrated in Fig. 30.9. Each block is connected to the loader plate by a spring with spring constant k_L. Adjacent blocks are connected to each other by springs with spring constant k_C.

The blocks interact with the surface through a static-dynamic friction law. The static stability of each slider-block is given by

$$k_L y_i + k_C (2y_i - y_{i-1} - y_{i+1}) < F_{Si} \tag{10}$$

where F_{Si} is the maximum static friction force on block i, and y_i is the position of block i relative to the loader plate.

When the cumulative force from the springs connected to block i exceeds the maximum static friction F_{Si}, the block begins to slide. The dynamic slip of block i is controlled by the equation

$$m\frac{d^2y_i}{dt^2} + k_L y_i + k_C\left(2y_i - y_{i-1} - y_{i+1}\right) = F_{Di} \tag{11}$$

where F_{Di} is the dynamic (sliding) frictional force on block i. The loader plate velocity is assumed to be much smaller than the slip velocity so that the movement of the loader plate is neglected during a slip event. The sliding of one block can trigger the instability of the other blocks forming a multiple block event.

It is convenient to introduce the nondimensional variables and parameters

$$\tau = t\sqrt{\frac{k_L}{m}},\ Y_i = \frac{k_L y_i}{F_S^{ref}},\ \phi = \frac{F_{Si}}{F_{Di}},\ \alpha = \frac{k_C}{k_L},\ \beta_i = \frac{F_{Si}}{F_S^{ref}} \tag{12}$$

The ratio of static to dynamic friction ϕ is assumed to be the same for all blocks but the values themselves β_i vary from block to block with F_S^{ref} as a reference value of the static frictional force.

In terms of these nondimensional variables the static stability condition Eq. 10 becomes

$$Y_i + \alpha\left(2Y_i - Y_{i-1} - Y_{i+1}\right) < \beta_i \tag{13}$$

and the dynamic slip Eq. 11 becomes

$$\frac{d^2Y_i}{d\tau^2} + Y_i + \alpha\left(2Y_i - Y_{i-1} - Y_{i+1}\right) = \frac{\beta_i}{\phi} \tag{14}$$

Before obtaining solutions, it is necessary to prescribe the parameters ϕ, α, and β_i. The parameter α is the stiffness of the system. We will give results for $\alpha = 3$ and $\alpha = 1000$. For $\alpha = 3$ the system is soft and there are no system wide (100 block) events. For $\alpha = 1000$ the system is stiff and system wide (100 block) events dominate. The ratio ϕ of static friction to dynamic friction is taken to be the same for all blocks $\phi = 1.5$, while the values of frictional parameters β_i are assigned to blocks with a uniform random distribution in the range $1 < \beta_i < 3.5$. This random variability in the system is a "noise" required to generate event variability in stiff systems.

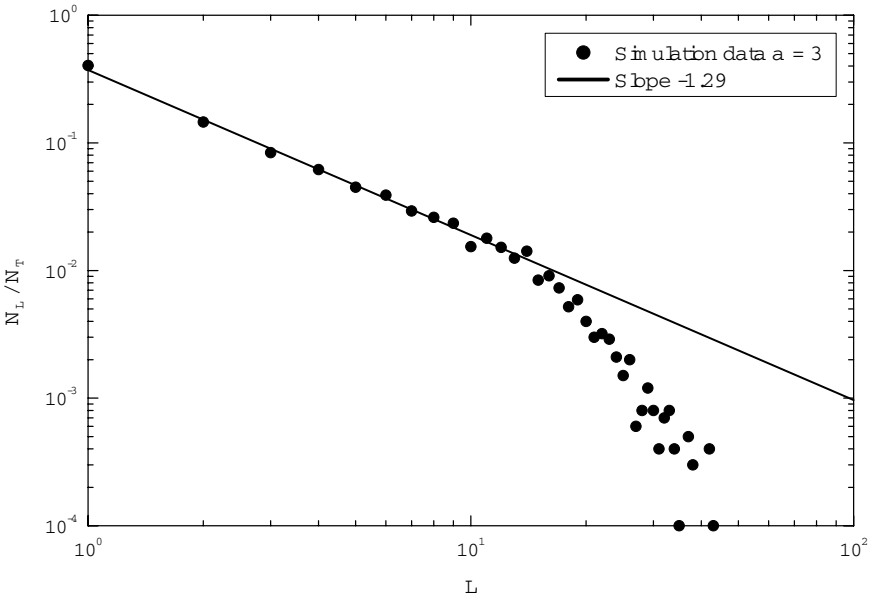


Fig. 30.10. Frequency-size distribution of 10,000 slip events for a "soft" system with $\alpha = 3$. The ratio of the number of events N_L of event size L to the total number of events N_T is given as a function of L. The solid line is a power-law dependence with exponent -1.29.

We first give results for a soft system with $\alpha = 3$. The linear event size L is the number of blocks that participate in the event. Frequency-size statistics for 10,000 events are given in Fig. 30.10. The smaller events are well approximated by a power-law relation with exponent -1.29. For larger event sizes there is a roll over and there are no system-wide (100 block) events. We next consider the interoccurrence time statistics for events. In order to do this we specify a threshold magnitude and consider events only larger than this threshold. The cumulative distribution $P(t)$ of interoccurrence times for events in the size range from $L = 10$ to 100 is given in Fig. 30.11. Also included in this figure is the best log likelihood (-1549.) fit of the exponential (Poisson) distribution to these data, this is Eq. 8 with $\gamma = 1$. To a good approximation the events occur randomly and are not correlated. A similar result was obtained for the sandpile model (Sanchez, Newman, and Carreras 2002). Since these smaller events occur at different positions along the array they do not have a memory of previous events and they occur randomly.

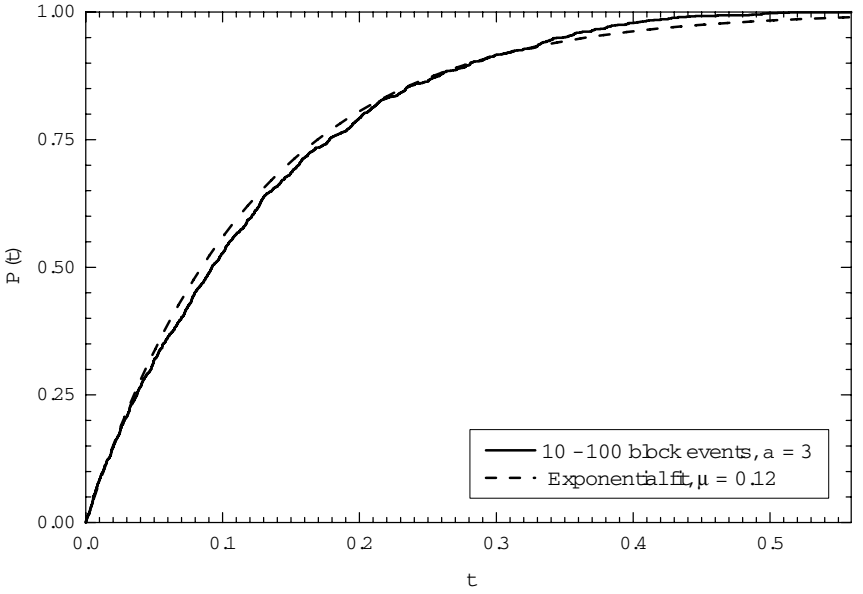


Fig. 30.11. Cumulative distribution function $P(t)$ of interoccurrence times t for the events given in Fig. 30.7 in the size range L = 10 to 100. The solid line is the distribution of observed recurrence times. The dashed line is the best-fit exponential (Poisson) distribution, Eq. (8) with $\beta = 1$.

We next give results for a stiff system with $\alpha = 1000$. The motion organizes itself into the recurrence of system-wide (100 block) events separated by sets of small size events. Frequency-size statistics for 10,000 events are given in Fig. 30.12. Again the smaller events are well approximated by a power-law relation with exponent -2.12. In this case there are about 1500 system-wide (100 block) events. We consider that these are equivalent to characteristic earthquakes. We next consider the recurrence time statistics for these events. The cumulative distribution of these recurrence times is given in Fig. 30.13. Also included in this figure is the best log likelihood (-1779.) fit of the Weibull distribution Eq. 8 to this data obtained by taking $\tau = 0.206 \pm 0.002$ and $\beta = 2.60 \pm 0.05$.

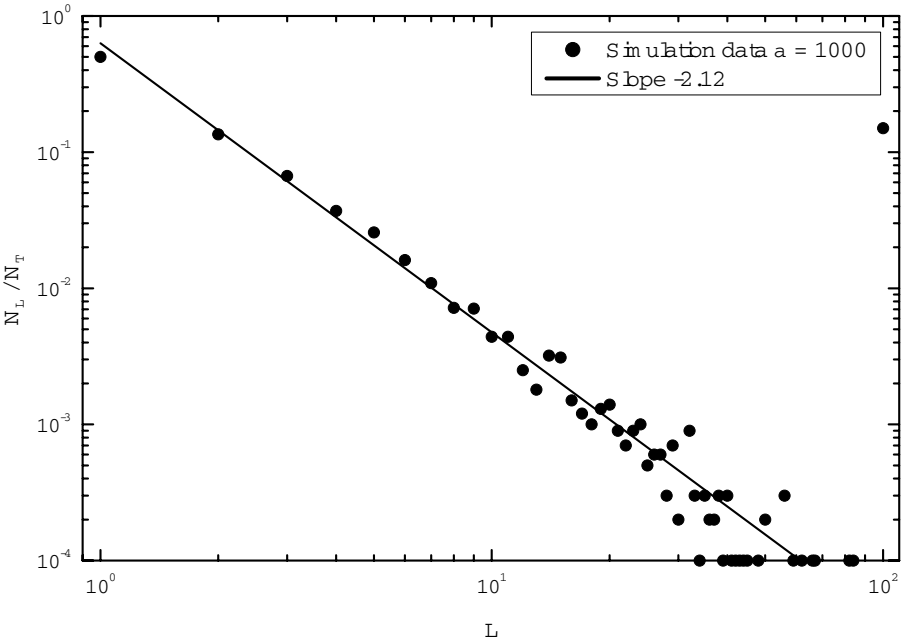


Fig. 30.12. Frequency-size distribution of 10,000 slip events for a "stiff" system with α=1000. The ratio of the number of events N_L of event size L to the total number of events N_T is given as a function of L. The solid line is a power-law dependence with exponent -2.12.

10 Gaussian Fractional Noises

A Gaussian fractional noise is a time series with a Gaussian distribution of values and a power-spectral density S that satisfies the condition

$$S(f) \propto f^{-\beta} \tag{15}$$

with $-1 < \beta < 1$ and f the frequency. With $\beta = 0$ this is a Gaussian white noise, with $0 < \beta < 1$ the time series is persistent (positive fallows positive, negative follows negative), and with $-1 < \beta < 0$ the time series is antipersistent (negative follows positive and positive follows negative). Gaussian white noises are stationary, the mean and variance are independent of the length of record T. For further details see Malamud and Turcotte (1999).

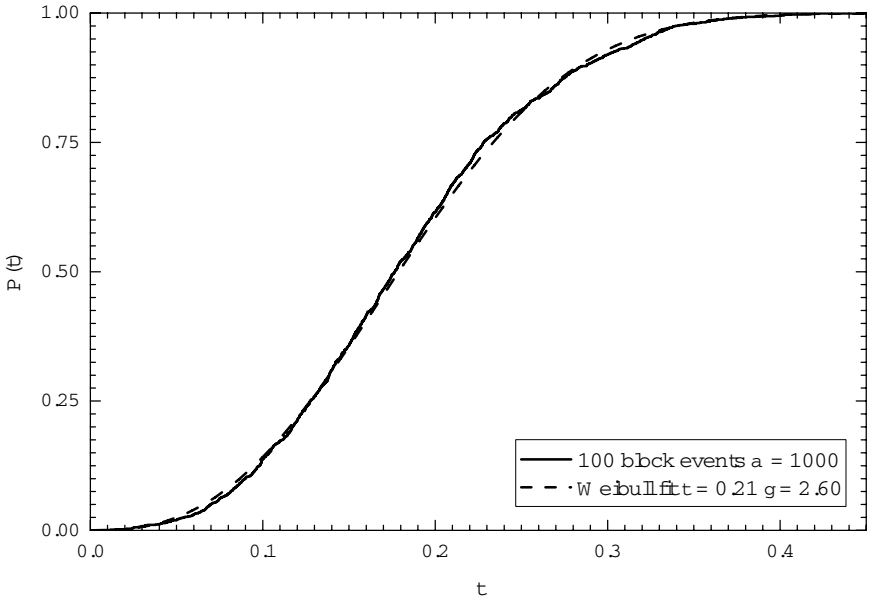


Fig. 30.13. Cumulative distribution function $P(t)$ of recurrence times t for the 1500 system-wide (100 block) events with $\alpha = 1000$. The solid line is the distribution of observed recurrence times. The dashed line is the best-fit Weibull distribution Eq.(8) with $\tau = 0.21$ and $\beta = 2.60$.

An alternative quantification of persistence and antipersistence in fractional noises is the autocorrelation function $C(s)$ where s is the lag. For a fractional Gaussian noise we have

$$C(s) \propto s^{-\gamma} \tag{16}$$

where

$$\gamma = 1\text{-}\beta \tag{17}$$

For a Gaussian white noise ($\beta = 0$) we have $\gamma = 1$, for a persistent fractional noise ($0 < \beta < 1$) we have $0 < \gamma < 1$, and for an antipersistent fractional noise ($-1 < \beta < 0$) we have $1 < \gamma < 2$.

Another alternative quantification of persistence and antipersistence in fractional noises is Hurst rescaled range analysis $R(T)$ where T is the length of record considered. For a fractional Gaussian noise we have

$$R(T) \propto T^{Hu} \tag{18}$$

where Hu is known as the Hurst exponent. The power-law exponents are related by

$$Hu = \frac{1+\beta}{2} = 1 - \frac{\gamma}{2} \tag{19}$$

For a Gaussian white noise ($\beta = 0$, $\gamma = 1$) we have $Hu = ½$, for a persistent fractional noise ($0 < \beta < 1$) we have $½ < Hu < 1$, and for an antipersistent fractional noise ($-1 < \beta < 0$) we have $0 < Hu < ½$.

Bunde, Eichner, Havlin, and Kantelhardt (2003) studied the distribution of interval times between peaks over threshold of fractional Gaussian noises. They determined the statistical distribution of the time intervals between the times when the fractional noise exceeded a specified threshold. They found that the cumulative distribution of interval times was well approximated by the Weibull distribution given in Eq. 8. The power-law exponent γ in the Weibull distribution was found to be equal to the power-law exponent γ in the autocorrelation function of the white noise as given in Eq. 16. For a Gaussian white noise ($\beta = 0$) with $\gamma = 1$ the Weibull distribution reduces to the Poisson distribution as expected. For persistent fractional noises ($0 < \beta < 1$) we have $0 < \gamma < 1$ and the Weibull distribution is a stretched exponential distribution. For antipersistent fractional noises ($-1 < \beta < 0$) we have $1 < \gamma < 2$. For a $1/f$ ($\beta = 1$) fractional noise the distribution of return times is well approximated by the Rayleigh distribution $\gamma = 2$. Other studies of the interval times of fractional noises have been carried out by Altmann, da Silva, and Caldas (2004); Altmann and Kantz (2005); Bunde, Eichner, Havlin, and Kantelhardt (2004); Eichner, Kantelhardt, Bunde, and Havlin (2006); Pennetta (2006).

11 Hazard Rate Versus Hazard Function

In Eq. 9 the hazard function was defined. In the engineering literature the fiber bundle and damage models are widely used to study the degradation and failure of materials (Krajcinovic 1996). The dynamic failure of a bundle of N_0 fibers is modeled empirically by the relation (Newman and Phoenix 2001)

$$\frac{dN}{dt} = -N\nu(\sigma) \tag{20}$$

where N is the number of unbroken fibers and $\nu(\sigma)$ is the hazard rate that is a function of the stress σ on the fibers.

The cumulative probability P that a fiber is broken is given by

$$P = 1 - \frac{N}{N_0} \tag{21}$$

And the pdf p for fiber failure is given by

$$p = \frac{dP}{dt} = -\frac{1}{N_0}\frac{dN}{dt} \tag{22}$$

From Eqs. 9, 20, 21, and 22 we obtain

$$h = -\frac{1}{N}\frac{dN}{dt} = \nu \tag{23}$$

Thus the hazard rate for fibers is equal to the hazard function for the distribution of fibers.

The standard definition of the damage variable α is given by (Kachanov 1986)

$$E_{eff} = E_0(1-\alpha) \tag{24}$$

where E_{eff} is the Young's modulus of the damaged material and E_0 is the Young's modulus of the undamaged ($\alpha = 0$) material. When $\alpha = 1$, failure occurs. For a fiber bundle the effective Young's modulus is given by

$$E_{eff} = (1 - \frac{N}{N_0})E_0 \tag{25}$$

Comparing Eqs. 24 and 25 we have

$$\alpha = 1 - \frac{N}{N_0} \tag{26}$$

relating the damage mechanics and fiber bundle models (Krajcinovic 1996).

It is standard practice in both the fiber bundle model and the damage mechanics model to assume that rates have a power-law dependence on stress. For the fiber-bundle model this is written (Newman and Phoenix 2001)

$$\nu(\sigma) = \nu_0 \left(\frac{\sigma}{\sigma_0} \right)^p \tag{27}$$

where ν_0 is the reference hazard rate at the reference stress σ_0.

Assuming that the stress increases linearly in time

$$\frac{\sigma}{\sigma_r} = \frac{t}{t_r} \tag{28}$$

Eqs. 20 and 27 become

$$\frac{dN}{dt} = -N\nu_r \left(\frac{t}{t_r} \right)^p \tag{29}$$

Integrating with $N = N_0$ at $t = 0$ gives

$$\frac{N}{N_0} = \exp\left[-\frac{\nu_r t^{p+1}}{p\tau^p} \right] \tag{30}$$

We next introduce

$$P(t) = 1 - \frac{N}{N_0}, \ \beta = p + 1, \ \nu_r = \frac{p}{\tau} \tag{31}$$

where $P(t)$ is the cumulative distribution function of failure. Substituting Eq. 31 into Eq. 30 gives the Weibull distribution Eq. 8. The self-similar (power-law) dependence of the hazard rate on time leads directly to the Weibull distribution of failure times.

The hazard function as given in Eq. 9 for the Weibull distribution has a power-law dependence on the time since the last event. The stress on a fault is expected to increase linearly in time due to the steady motion of the tectonic plates just as the forces on our slider blocks increase linearly in time due to the steady motion of the loader plate. Similarly the stress on the fiber bundle as given by Eq. 28 increases linearly with time. We have shown above that the failure rate of a fiber bundle

modeled by Eq. 20 gives a Weibull distribution. This provides a basis for the application of the Weibull distribution both to our slider-block model and to characteristic earthquakes.

Acknowledgment

This work has been supported by the National Science Foundation Grant ATM-0327571. We would like to acknowledge the contributions of Bruce Malamud and an anonymous reviewer to the improvement of this manuscript.

References

Altmann, E.G., da Silva, E.C. and Caldas, I.L. (2004) Recurrence time statistics for finite size intervals. Chaos 14(4), 975-981.

Altmann, E.G. and Kantz, H. (2005) Recurrence time analysis, long-term correlations, and extreme events. Phys. Rev. E 71(5), 056106.

Bak, P., Tang, C. and Wiesenfeld, K. (1988) Self-organized criticality. Phys. Rev. A 38(1), 364-374.

Biasi, G.P., Weldon, R.J., Fumal, T.E. and Seitz, G.G. (2002) Paleoseismic event dating and the conditional probability of large earthquakes on the southern San Andreas fault, California. Bull. Seism. Soc. Am. 92(7), 2761-2781.

Bucknam, R.C., Coe, J.A., Chavarria, M.M., Godt, J.W., Tarr, A.C., et al. (2001) Landslides triggered by Hurricane Mitch in Guatemala - inventory and discussion, U.S. Geological Survey (Open-File Report No. 2001-443).

Bunde, A., Eichner, J.F., Havlin, S. and Kantelhardt, J.W. (2003) The effect of long-term correlations on the return periods of rare events. Physica A 330(1-2), 1-7.

Bunde, A., Eichner, J.F., Havlin, S. and Kantelhardt, J.W. (2004) Return intervals of rare events in records with long-term persistence. Physica A 342(1-2), 308-314.

Burridge, R. and Knopoff, L. (1967) Model and theoretical seismicity. Bull. Seism. Soc. Am. 57, 341-371.

Carlson, J.M. and Langer, J.S. (1989) Mechanical model of an earthquake fault. Phys. Rev. A 40(11), 6470-6484.

Drossel, B. and Schwabl, F. (1992) Self-organized critical forest-fire model. Phys. Rev. Lett. 69(11), 1629-1632.

Eichner, J.F., Kantelhardt, J.W., Bunde, A. and Havlin, S. (2006) Extreme value statistics in records with long-term persistence. Phys. Rev. E 73(1).

Gabrielov, A., Newman, W.I. and Turcotte, D.L. (1999) Exactly soluble hierarchical clustering model: Inverse cascades, self-similarity, and scaling. Phys. Rev. E 60(5), 5293-5300.

Guzzetti, F., Malamud, B.D., Turcotte, D.L. and Reichenbach, P. (2002) Power-law correlations of landslide areas in central Italy. Earth Planet. Sci. Lett. 195(3-4), 169-183.

Harp, E.L. and Jibson, R.L. (1995) Inventory of landslides triggered by the 1994 Northridge, California earthquake, U.S. Geological Survey (Open-File Report No. 1995-213).

Kachanov, L.M. (1986) *Introduction to continuum damage mechanics.* Martinus Nijhoff, Dordrecht.

Korvin, G. (1992) *Fractal models in the Earth sciences.* Elsevier, Amsterdam.

Krajcinovic, D. (1996) *Damage mechanics.* Elsevier, Amsterdam.

Malamud, B.D., Millington, J.D.A. and Perry, G.L.W. (2005) Characterizing wildfire regimes in the United States. Proc. Natl. Acad. Sci. U. S. A. 102(13), 4694-4699.

Malamud, B.D., Morein, G. and Turcotte, D.L. (1998) Forest fires: An example of self-organized critical behavior. Science 281(5384), 1840-1842.

Malamud, B.D. and Turcotte, D.L. (1999) Self-affine time series: I. Generation and analyses. Adv. Geophys. 40, 1-90.

Malamud, B.D. and Turcotte, D.L. (2000) Cellular-automata models applied to natural hazards. Computing in Science & Engineering 2(3), 42-51.

Malamud, B.D., Turcotte, D.L., Guzzetti, F. and Reichenbach, P. (2004) Landslide inventories and their statistical properties. Earth Surface Processes and Landforms 29(6), 687-711.

Mandelbrot, B. (1967) How long is coast of Britain - Statistical self-similarity and fractional dimension. Science 156(3775), 636.

Meeker, W.Q. and Escobar, L.A. (1991) *Statistical methods for reliability data.* John Wiley, New York.

Millington, J.D.A., Perry, G.L.W. and Malamud, B.D. (2006). Models, data and mechanisms: quantifying wild wildfire regimes. In G. Cello & B.D. Malamud (Eds.), *Fractal analysis for natural hazards* (pp. 155-167). London: Geological Society.

Minnich, R.A. (2001) An integrated model of two fire regimes. Conservation Biology 15(6), 1549-1553.

Newman, W.I. and Phoenix, S.L. (2001) Time-dependent fiber bundles with local load sharing. Phys. Rev. E 63(2), 021507.

Niklasson, M. and Granstrom, A. (2000) Numbers and sizes of fires: Long-term spatially explicit fire history in a Swedish boreal landscape. Ecology 81(6), 1484-1499.

Patel, J.K., Kapadia, C.H. and Owen, D.B. (1976) *Handbook of statistical distributions.* Marcel Dekker, New York.

Pennetta, C. (2006) Distribution of return intervals of extreme events. Eur. Phys. J. B 50(1-2), 95-98.

Ricotta, C., Arianoutsou, M., Diaz-Delgado, R., Duguy, B., Lloret, F., et al. (2001) Self-organized criticality of wildfires ecologically revisited. Ecological Modelling 141(1-3), 307-311.

Rundle, J.B., Rundle, P.B., Donnellan, A. and Fox, G. (2004) Gutenberg-Richter statistics in topologically realistic system-level earthquake stress-evolution simulations. Earth Planets Space 56(8), 761-771.

Rundle, J.B., Rundle, P.B., Donnellan, A., Turcotte, D.L., Shcherbakov, R., et al. (2005) A simulation-based approach to forecasting the next great San Francisco earthquake. Proc. Natl. Acad. Sci. U. S. A. 102(43), 15363-15367.

Rundle, J.B., Turcotte, D.L., Shcherbakov, R., Klein, W. and Sammis, C. (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. Rev. Geophys. 41(4), 1019.

Sanchez, R., Newman, D.E. and Carreras, B.A. (2002) Waiting-time statistics of self-organized-criticality systems. Phys. Rev. Lett. 88(6), 068302.

Savage, J.C. (1994) Empirical earthquake probabilities from observed recurrence intervals. Bull. Seism. Soc. Am. 84(1), 219-221.

Turcotte, D.L. (1997) *Fractals and chaos in geology and geophysics.* (2nd ed.)Cambridge Univ. Press, Cambridge.

Turcotte, D.L. (1999) Self-organized criticality. Rep. Prog. Phys. 62(10), 1377-1429.

Weibull, W. (1951) A statistical distribution function of wide applicability. Journal of Applied Mechanics-Transactions of the Asme 18(3), 293-297.

Yakovlev, G., Newman, W.I., Turcotte, D.L. and Gabrielov, A. (2005) An inverse cascade model for self-organized complexity and natural hazards. Geophys. J. Int. 163(2), 433-442.

Yakovlev, G., Turcotte, D.L., Rundle, J.B. and Rundle, P.B. (2006) Simulation based distributions of earthquake recurrence times on the San Andreas fault system. Bull. Seism. Soc. Am. 96, 1995-2007.

31 Predicting the Multifractal Geomagnetic Field

D. Vassiliadis,[1] A. Pulkkinen,[2] A.J. Klimas,[3] D.N. Baker[4]

1 ST at NASA/GSFC and George Mason University, dvassili@pop600.gsfc.nasa.gov
2 UMBC/GEST at NASA/GSFC, antti.pulkkinen@nasa.gov
3 NASA/Goddard Space Flight Center, alex.klimas@nasa.gov
4 University of Colorado at Boulder/LASP, daniel.baker@lasp.colorado.edu

Abstract. The geomagnetic field at the Earth's surface is determined by a large number of current sources, internal and external, which vary at overlapping time scales. As a result, the time variation of the field, ranging from nHz to tens of Hz, can be modeled as a multifractal whose properties depend on frequency and activity level, with the latter being largely determined by the external driving due to the interplanetary medium. Multifractal properties are apparent in all measures, conventional or modern, of magnetic activity: the magnitude and individual components; regional and global nonlinear averages known as magnetic indices; and nonstationary oscillations due to wave excitation and damping. In recent years, we have developed predictive models of the three types of measures which are representations of the field's modes of response to interplanetary magnetic and plasma variables. Thus, although the field's time variation appears multifractal, many of its features are predictable since they are related to nonlinear modes in response to the turbulent interplanetary medium. We present two examples: the predictive modeling of the high-latitude field disturbances as a response to the interplanetary electric field, and that of its derivative, $\partial \mathbf{B} / \partial t$, at mid- and high latitudes as a response to the solar wind plasma moments. We discuss the merits and disadvantages of the "modal" versus the "multifractal" approach.

1 Introduction

Temporal variations in the magnetic field at the surface of the Earth and in the surrounding geospace have been known to occur over a wide variety of scales ranging from fractions of seconds for low-frequency waves to many years for the solar-cycle and secular variations. The power spectrum of these variations generally falls off as ω^{-k} where the exponent k varies between -1 and -2 depending on the frequency range and the activity level. The few exceptions to the power law are the diurnal variation peak and, at frequencies higher than 1 mHz, field-line resonances. Going beyond the power spectrum, in the last 20 years there has been additional evidence from correlation function and structure function analysis that the magnetic field is a multifractal, varying at all observable temporal and spatial scales, and featuring

events that have a finite occurrence probability. Here we adopt the standard definition of a multifractal, namely that it is an object which is statistically invariant under a change of (spatial/temporal) scale although the scaling exponent slowly changes from one scale range to another.

At the same time it has been recognized that nonlinear models can be used to deterministically reproduce and even predict the main types of events that appear in time series of various magnetic-field measures. Geomagnetic disturbances, known since the late 17^{th} century, comprise diverse events as global as magnetic storms and magnetospheric substorms and as transient as MHD wave excitation. The first few years of spaceflight led to the experimental confirmation that geomagnetic disturbances on the ground are in fact consequences of global (planetary) geospace events. The statistical recurrence of these geospace events is sufficiently significant that it can serve as the basis of nonlinear dynamic models that are used to predict their ground-magnetic signatures to a good accuracy.

In the last 20 years, therefore, different methods of nonlinear dynamics have been applied to reach two apparently contradictory conclusions: the multifractal, or self-similar, character of the field over wide frequency or energy ranges (according to some viewpoints, this experimental fact casts the field as stochastic and difficult to predict); and modeling of several types of geospace events, reflected in the field variability, based on nonlinear models. Several papers have discussed the "dual" nature of magnetic field disturbances, such as "multifractal" and "global" and presented different ways to describe and reconcile the two viewpoints.

In this paper we review the two approaches and present their respective strengths and limitations. In some regimes the multifractal approach is particularly useful in demonstrating the effect of the external driver(s) of geomagnetic activity, namely the interplanetary plasma of the solar wind and its field, the interplanetary magnetic field (IMF). Multifractal properties of the solar wind and IMF have been measured for a long time [Burlaga, 1995]. In other regimes, multifractal scalings are evident, but are not related to interplanetary features and must be associated with certain stochastic magnetospheric properties. On the other hand, dynamic models represent the complex, driven, dissipative, large-scale (and mostly magnetofluid than kinetic) processes in the magnetosphere and ionosphere.

The present debate is not unlike an earlier, by now resolved controversy regarding a) the nonlinearity and b) the low-dimensionality of geomagnetic variations. Correlation dimension analysis indicated that the dimension d_2 was of the order of 2-4 for AE and other electrojet indices [Vassiliadis et al., 1990; Roberts et al., 1991; Shan et al., 1991]. In the early 1990s these results were interpreted as evidence for low dimension and nonlinearity. However, it was pointed out that correlation dimension analysis was not appropriate for geomagnetic variations because, as will be discussed below, they are strongly driven by the interplanetary medium [Prichard and Price, 1993]. Fractal-dimension analysis is better suited for autonomous (non-interacting) systems, while for driven systems is it useful mainly as a diagnostic of the power spectrum [Osborne and Provenzale, 1989]. In addition, technical points were raised [Price and Prichard, 1992]. Subsequently, several papers showed that input-output nonlinear dynamics models were able to represent different aspects of the geomagnetic variations while being low-dimensional either explicitly [e.g., Price

et al., 1994; Klimas et al., 1994; Vassiliadis et al., 1995, 1996a,b; Horton and Doxas, 1996] or implicitly [e.g., Gleisner and Lundstedt, 1997]. Thus nonlinearity was shown to be generally valid, while low-dimensionality was evident in at least several intervals. In addition, independent studies using autonomous methods and addressing the technical points raised earlier, showed that d_2 seems to be low in several intervals of interest [Pavlos et al., 1999a,b]. At the conclusions section of this chapter we discuss how the part of that earlier debate regarding low-dimensionality is resurrected in the current "modal versus multifractal" debate.

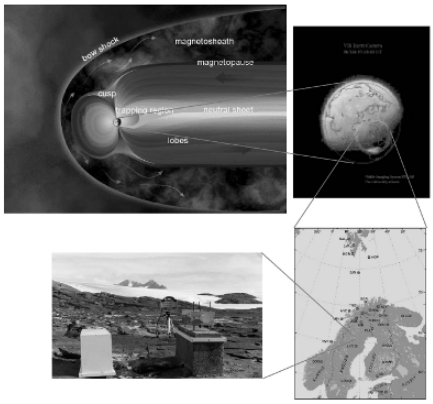


Fig. 31.1. Sources of the geospace magnetic field: crustal, magnetospheric, ionospheric, and induced currents flow in the geospace and are measured by magnetometers which are deployed either individually or as part of an array. (A color version of this figure appears between pages 196 and 197).

2 Temporal and Spatial Variability of Geospace Currents

The field at the Earth's surface and in the volume of the magnetosphere is determined by primary currents flowing in the magnetosphere and ionosphere, and secondary currents induced in the planetary crust. The most important currents flowing in geospace are the ring current, the tail current, and, in the polar/auroral regions, the auroral electrojets and field-aligned currents (Fig. 31.1). These and other currents are produced by processes such as magnetic reconnection and magnetic flux transfer in the magnetosphere, electrodynamics in the ionosphere and the magnetospheric plasma sheet, and viscous processes in boundary layers. Each current does not flow in isolation, but they interconnect in several different ways, depending on the activity state of geospace, to form regional and global current systems.

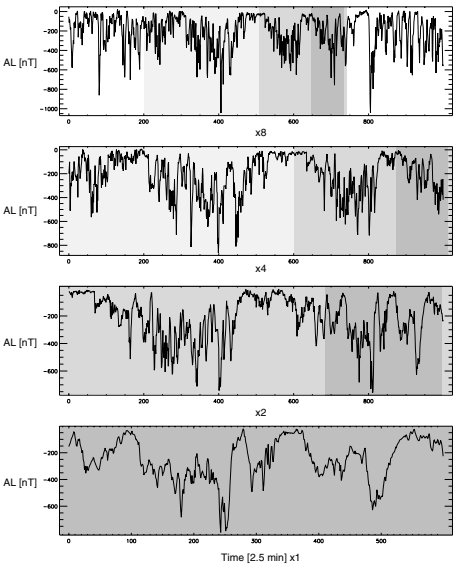


Fig. 31.2. Statistical self-similarity of the geomagnetic index AL. A segment of the Bargatze et al. [1985] dataset is shown in increasing magnification as indicated by the shaded intervals.

The currents fluctuate in magnitude and position over a wide range of time scales, from fractions of seconds to tens of years. Thus the variability of the field is a representation of the interplanetary variability as well as nonlinearities and instabilities internal to the magnetosphere and ionosphere. There is significant evidence for the fractal (self-similar) character of the field, from several statistical studies and it is evident from invariance under change of scale (as illustrated in Fig. 31.2).

There is less evidence for the spatial variability of the field, but it can be surmised from measurements of related variables. Based on the present configuration of magnetometer arrays, the surface field is found to be correlated over large local-time ranges, comparable to the planetary range, and smaller latitudinal ranges. The large-scale local-time correlation is the net effect of particle drifts and current systems that are organized zonally. The spatial extent of these current/magnetic structures ranges from tens and hundreds of kilometers (surface anomalies; ionospheric structures; field-aligned structures) to thousands of kilometers (large-scale ionospheric currents) to tens of Earth radii (magnetospheric currents). The sparseness of the magnetometers in a typical array ($\geq$100 km) is a limitation for quantifying the intrinsic spatial variability [Pulkkinen et al., 2006]. However, remote-sensing of the magnetosphere and ionosphere (in various ranges such as radar, optical, and UV) reveals a high variability in plasma structure and dynamics, and therefore suggests similarly complex current and field structure.

A defining factor for the spatial and temporal complexity of the field is its reconfiguration during events of manifestly nonlinear magnetospheric and ionospheric response, as will be discussed in more detail in Section 5 below. Arguably the most characteristic storage-release magnetospheric process is the global instability called a substorm. During such an event, the stored magnetospheric energy is released, and partly dissipated, via the process of collisionless magnetic reconnection. The magnetospheric topology changes dramatically: one or more segments of the nightside magnetosphere, or magnetotail, the so-called plasmoids, are disconnected, and travel antisunward until they are eventually swept by the solar wind. These large-scale changes have complex geomagnetic signatures in space and on the ground.

In summary, Earth's magnetic field has significant spatial and temporal variability, partly arising from intrinsic instability and partly from the interaction with the interplanetary environment. The next section discusses measures that have been developed to represent this variability and the following one, how this complexity is described in a multifractal framework.

3 Measures of the Field

We saw that the variation of the magnetospheric magnetic field $\mathbf{B}(\mathbf{x};t)$ is determined by a large number of global and local, electrodynamic and plasma physical processes. The field is not always the best measure of the amplitude of all these processes; secondary (derived) variables need to be derived in such a way that they can manifest different properties and the effects of different processes. The usefulness of each variable depends on the process of interest, its amplitude in the measurement

location in geospace (incl. the Earth's surface), on time (seasonal, diurnal, etc.), and on other user-defined issues.

We therefore distinguish between three measures of magnetic activity as follows:

1. The local field **B**(**x**,t) at location **x** and time t is the primary variable from which the others are derived. In this paper we pay particular attention to the surface field, but many conclusions can be generalized to the magnetospheric field, and even to the geospace field (which is also measured in the interplanetary medium). The variable of interest is the disturbed field,

$$\delta\mathbf{B}\left(\mathbf{x},t\right) \equiv \mathbf{B} - \mathbf{B}_q$$

where $\mathbf{B}_q$ is the contribution of the diurnal ionospheric quiet current S_q and slower-varying sources. Because of the dipole orientation, one is typically interested in the North-South field component which responds to currents flowing mainly in the east-west direction such as the auroral electrojets, the ring current, tail current, magnetopause current, etc. We will use B(**x**,t) to denote the disturbed-field component.

2. There are problems, however, with extracting significant amounts of physical information directly from one or few locations. Field measurements contain the effects of various current sources even under quiet conditions. Permanent ground magnetometers have traditionally been spaced at distances much larger than 1000 km; more recently magnetometer arrays are developed with distances between adjacent magnetometers of at least 100 km and spanning distances of thousands of km. To obtain useful information from sparse measurements the geomagnetic community has adopted geomagnetic indices of activity, which are nonlinear averages or envelopes of the magnetic disturbance in a region of interest [Mayaud, 1980; Menvielle and Berthelier, 1991]. These regional or global diagnostics are designed to be more robust than measurements of the local field.

Some of the best-known indices are the auroral electrojet indices, AL, AU, and AE, derived from local field measurements in the auroral zone, a region of high magnetic and auroral (UV, optical) activity. The AL index measures the strongest Southward deflection (due to a Westward current) and the AU measures the strongest Northward deflection (due to an Eastward current) in that region. Their difference

$$AE = AU - AL$$

is a measure of the total disturbed field in the auroral zone. Typical measurements of the indices have a 1-min or lower time resolution. Other geomagnetic indices include D_{st} which is a measure of the mid-latitude disturbance produced

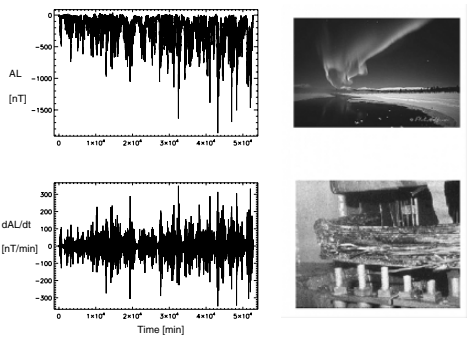


Fig. 31.3. Two different measures of geomagnetic activity: the AL index linked to ionospheric currents and auroral displays; and its time derivative, dAL/dt, linked to geoelectric currents and damaging effects of space weather on ground structures such as electric power grid components.

by the ring current, K_p and A_p which are measures of planetary activity, and PC which is a measure of the polar-region magnetic disturbance.

3. An important part of the geomagnetic activity is due to the secondary currents flowing in the Earth's crust [Viljanen et al., 1999]. Ionospheric and field-aligned currents induce geoelectric fields. Therefore the time derivative of the field, $\partial \mathbf{B}\left(\mathbf{x},t\right)/\partial t$, or a regional index (e.g., dAE/dt) has been recently identified as a distinct and important measure of geomagnetic fluctuations in its own right. The time derivative and related functions (such as $\left|\partial/\partial t\right|^2$) measure the amplitude, or power, of magnetic fluctuations around a slower-varying trend or oscillation.

Thus each measure of magnetic field activity (local, regional average, time derivative) measures different properties of the magnetospheric activity. The sample probability distribution and the power spectrum have different properties. Moreover, if the variables are obtained from the field through a nonlinear transformation, the multifractal properties of the new variables and the original field are generally significantly different.

The variables also differ in their usage in space physics (Fig. 31.3): Traditionally, in space physics indices have been important in quantifying the (average) geospace conditions surrounding an event. With extended magnetometer arrays and proper visualization tools, direct measurements of the field disturbance are increasingly used, although indices are still important in summarizing the activity level. In the advent of space weather studies, the time derivative $\partial \mathbf{B}/\partial t$ has been recognized as a significant variable since it drives geoelectric fields and thereby geomagnetically induced currents.

4 Multifractal Properties of Magnetic-Field Measures

The physics of self-similarity is the invariance of a process with scale (for instance: length, time, energy, etc.) and, as a result, the scaling of an observable as a power law of that scale. While power laws are ubiquitous in nature, there are instances where the power law is only apparent and in reality a small part of a non-scalefree feature. A rule of thumb is that the scaling range should span at least 3 orders of magnitude. In addition, the physical interpretation of a power law scaling is not straightforward for most situations: there are many statistical processes that produce power laws [Sornette, 2004].

In the case of the magnetosphere, such instances of scale invariance seem to characterize the fundamental interaction of elementary (local) currents and fields. The invariance is generally different for different measures and different scales, therefore the self-similar field is a multifractal rather than a monofractal.

The temporal self-similarity of geomagnetic activity was documented before the application of nonlinear dynamics methods. Several studies showed that the power spectra of geomagnetic fluctuations were of power-law form with exponents close to 1.5. For instance, the self-affinity of AE has a Hoelder exponent of H=0.5 up to a time of 113 min. This timescale may be related to the substorms duration of 2-4

hours. Similar results for the self-affinity have been found for time series of field component [Takalo, 1993] and for the structure function both in time and space [Pulkkinen et al., 2006a].

Examined in more detail, however, geomagnetic variations do not exhibit a simple power-law power spectrum. A seminal study by Tsurutani et al. [1990] established that the geomagnetic AE index has in fact a double-power-law distribution with exponents -1.02 for $f < (5h)^{-1}$ and -2.42 for $f > (5h)^{-1}$. The 5-hour break was interpreted as the effect of substorms which last 2-4 hours [Lui, 1991] or recur on average every ~2.75 hours [Borovsky et al., 1993], and the higher-frequency part of power spectrum indicated a fast decrease of power, interpreted as rapid dissipation. We discuss the Tsurutani et al. [1990] result in some more detail below.

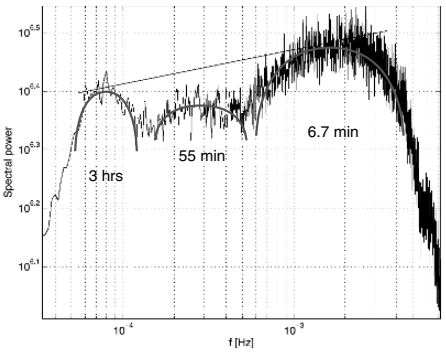


Fig. 31.4. The power spectrum of dAE/dt (after [Pulkkinen et al., 2006]).

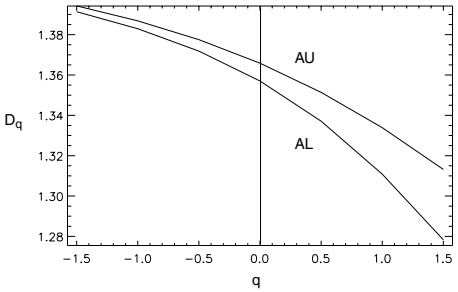


Fig. 31.5. The Renyi dimension spectrum of the AL and AU geomagnetic indices from the 2nd half of the Bargatze et al. [1985] dataset.

More recently, a structure function analysis for the AE index showed that it deviates significantly from a Markov process [Pulkkinen et al., 2006b]. In the same study, the power spectrum of the time derivative dAE/dt was analyzed and evidence was provided for three regimes (Fig. 31.4):

- Fast variations, ranging from 4 to 20 min peaking at 6.7 min. This peak carries most of the power, and include direct driving by the solar wind, as well as low-frequency MHD waves.
- Intermediate variations, ranging from 25 to 110 min peaking at 55 min. These are clustered around "loading-unloading" timescales, or those related to storage of interplanetary energy in the magnetotail and delayed release.
- Slow variations, ranging from 2 to 5.6 hours peaking at 3 hours. This peak is lowest in power and reflects slow driving by the solar wind and IMF.

While the peak power is not very different for these three regimes, the structuring is fairly evident and shows that there are specific "preferred" modes of response to the solar wind input whose derivative does not show significant peaks at those times.

In fact a multi-peak power spectrum is well suited to explain the complexity of geomagnetic variations of indices and fields. We have computed the Renyi dimensions

$$D_q = \lim_{\varepsilon \to 0} \frac{\frac{1}{1-q} \log \sum_i p_i^q}{-\log \varepsilon} \qquad (1)$$

for the AL and AU geomagnetic indices and find a significant variation with q (Fig. 31.5), indicating a multifractal, with the variation being greater for AL. This result is consistent with earlier computations of the D_q spectrum using a partition function

$$\Gamma(q,\tau) = \sum_{\Lambda_i} p_i(\tau)^q \approx \tau^{(q-1)D_q} \qquad (2)$$

where p_i is based on a positive-definite measure of the electrojet index, in this case AE [Consolini et al., 1991]. However, it should be noted that the index time series are assumed stationary while in reality they are not [e.g., Prichard and Price, 1993]. Also both the partition function and the Renyi dimension spectrum do not account for the strong effects of the intermittent interplanetary driver.

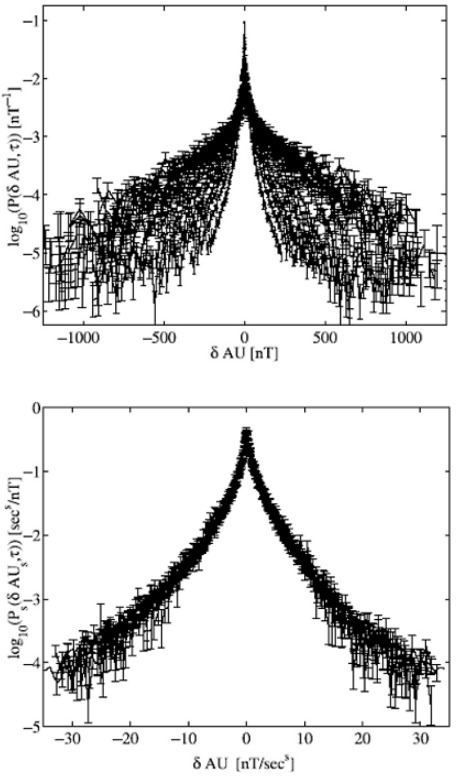


Fig. 31.6. PDF of time-differenced AU index: a. before and b. after finite-size scaling (from [Hnat et al., 2002]).

An attempt to correct the neglect of the solar wind input was made in a series of finite-size scaling studies [Hnat et al., 2002]. A Fokker-Planck model was used to describe the geomagnetic variations, which were considered as "statistical fluctuations" while interplanetary driving was replaced by low-amplitude noise.

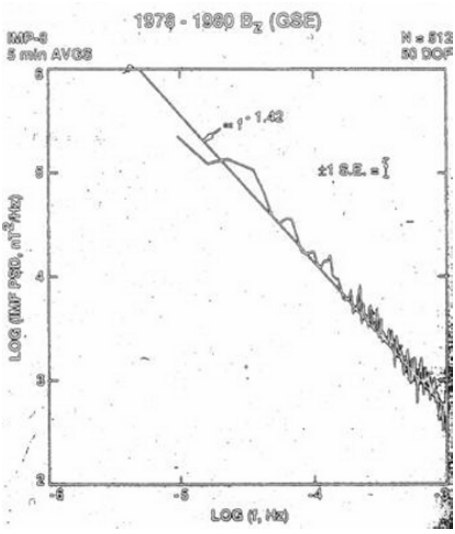
$$\frac{\partial P(\Delta B,t)}{\partial t} = \nabla\left(\gamma(\Delta B)\cdot P\right) + \nabla^2\left[D(\Delta B)\cdot P\right] \qquad (3)$$

where ΔB is a scalar measure of the magnetic field fluctuations $\Delta B(t,\tau) = B(t+\tau) - B(t)$. In different studies B represents an auroral electrojet index such as AE [Hnat et al., 2002] or a component of a local measurement [Pulkkinen et al., 2006]. In Eq. (3), P(B,t) is a probability density function (PDF) of ΔB, while γ(ΔB) and D(ΔB) are friction and diffusion coefficients. The probability measure is typically non-Gaussian and is interpreted as quantifying the "intermittency" of ΔB(t).

For certain choices of D(B) the probability has a finite-size scaling form:

$$P\left(\Delta B,\tau\right)=\tau^{-s}P_s\left(\Delta B\cdot\tau^{-s}\right)$$

where the first term on the right gives the self-similarity and the second is a finite-size scaling or "cutoff" term (often modeled as an exponential). For the geomagnetic indices AE, AL, and AU as well as for the solar wind driver $\varepsilon\equiv l_{ms}^2 v_{sw}B^2/\mu\cdot\sin\left(\theta/2\right)$, the PDF clearly shows finite-size scaling (Fig. 31.6). Similar finite-size scaling has been found for auroral emission UV-intensity variations which reflect the unloading behavior [Uritsky et al., 2006].



5 External Driving and Modes of Response

Geomagnetic variations across different length, time, and energy scales provide us with information for how large-scale order emerges in geospace plasmas and how it is eroded. In the spatial domain, the large-scale structure is determined by the internal field, the average solar wind pressure, and the patterns of solar illumination. The order in the temporal domain on the other hand, i.e., the appearance of disturbances such as storms, substorms, etc., is often a result of nonlinear, storage-release processes in the magnetosphere and ionosphere. Thus, while the drivers of geospace events are time-dependent, multifractal interplanetary field and plasma inputs, the final outcome is in addition determined by the nonlinear response.

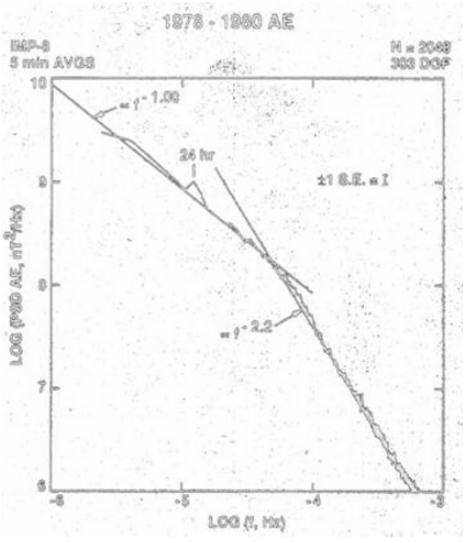


Fig. 31.7. Power spectra of solar wind $v_{sw}B_s$ and AE (from [Tsurutani et al., 1990]).

The interplanetary medium is a turbulent plasma with superimposed transient structures such as interplanetary shocks and magnetic clouds. However, these structures are relatively rare and the residual turbulent fields and plasma moments have

self-similar distributions [Burlaga, 1995]. The solar wind and the IMF interact with the magnetosphere and transfer energy, momentum, and mass to it. The efficiency of the coupling depends sensitively on the variables. For instance, magnetic reconnection between the IMF and the magnetospheric field is most efficient on the dayside when the fields are antiparallel, i.e., when the IMF B_z component is negative. In fact a simple representation of the energy transfer rate during reconnection involves the "Southward" field component

$$B_s = \begin{cases} 0, & B_z \geq 0 \\ -B_z, & B_z < 0 \end{cases} \qquad (4)$$

multiplied by the solar wind velocity, together producing the interplanetary electric field (IEF) $v_{sw}B_s$. The IEF is a measure of the energy transferred to the magnetosphere due to magnetic reconnection in the subsolar dayside region. Further, the coupling occurs at a continuum of time scales, ranging from minutes (passage of an interplanetary shock) to several days (passage of a cloud or high-speed stream) to the solar-rotation and solar-cycle variations. Similarly, the coupling may be limited to a narrow region in the dayside (cusp or subsolar point) or involve the entire magnetosphere/ionosphere (compression).

Because of the nonlinear, time-dependent coupling, the distributions and spectra of geomagnetic disturbance amplitudes differs from those of solar wind disturbances. Returning to the power-spectral analysis by [Tsurutani et al., 1990] of the geomagnetic index AE and the solar wind $v_{sw}B_s$, at low frequencies ($<(5\ \text{hours})^{-1}$) the spectra of both variables are power laws with exponents -1.00 and -1.42, respectively. Beyond 5 hours, the AE spectrum shows a change in slope to -2.2 as discussed above, whereas the $v_{sw}B_s$ spectrum remains unchanged (Fig. 31.7). Thus at high frequencies there is rapid dissipation of the energy transferred to the magnetosphere by the solar wind, and this dissipation mechanism involves substorms and similar relaxation mechanisms.

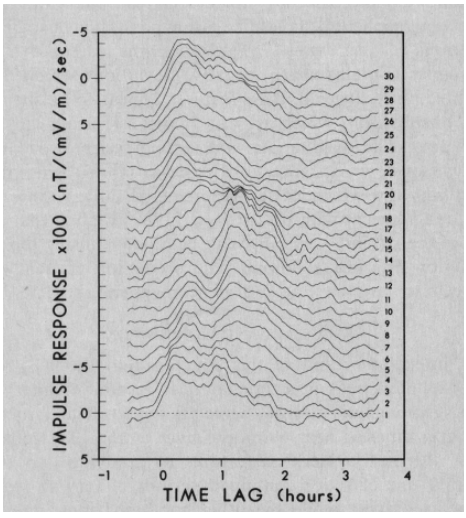


Fig. 31.8. Impulse response functions of AL to $v_{sw}B_s$, stacked in order of increasing activity level (from [Bargatze et al., 1985]).

Similar results were obtained from different studies as well. An event-statistical analysis of AE and the IEF measure $v_{sw}B_T \sin^4(\theta/2)$ estimated the dissipation range to start at 3.5 hours [Uritsky et al., 2001a]. And Langevin models developed from fluctuations of the AE index

and the Poynting flux in the solar wind clearly showed evidence for non-Markovian dynamics for the former, but not for the latter [Pulkkinen et al., 2006b].

The above results seem at odds with those from the finite-size scaling analysis of time-differenced auroral electrojet indices and the ε parameter, which show that the time differenced variables share nearly identical scaling parameters and very similar rescaled PDFs [Hnat et al., 2002]. The agreement is striking, but we note that the PDFs are dominated by the low-amplitude differences and break away from self-similarity, and from each other, at the more interesting high-amplitude changes (such as those that would characterize substorm expansion). The similarity among rescaled PDFs is also notable, but is largely dependent on the characteristics of the distribution near ΔB=0. Probably more interesting are the large-scale changes that occur at the wings of the distribution. In addition, the notion of the fluctuation, which underlies the approach, is not well justified for the time differences. Thus the geomagnetic output, while multifractal, possesses different properties than its interplanetary driver. To understand how the properties arise dynamically, we review a simple model of the coupling such as the linear finite-impulse response (FIR) model between the AL index and the IEF $v_{sw}B_s$:

$$AL(t) = \int_0^\infty H^{AL}_{v_{sw}B_s}(\tau)\, v_{sw}(t-\tau)\, B_s(t-\tau)\, d\tau \qquad (5)$$

The impulse response function $H^{AL}_{v_{sw}B_s}(\tau)$ is computed from historic time series data and constitutes a general, linear, time-invariant model for the dynamics of AL.

A linear model represents only the qualitative geomagnetic variation, and it can be generalized to a nonlinear one, such as the activity-parametrized FIR model [Bargatze et al., 1985; Clauer et al., 1986]:

$$AL(t) = \int_0^\infty H^{AL}_{v_{sw}B_s}(\tau; \overline{AL})\, v_{sw}(t-\tau)\, B_s(t-\tau)\, d\tau \quad (6)$$

where the impulse response depends on a measure of the AL activity level, here the sample mean over a given interval. As the AL activity level increases, there are qualitative changes in the impulse response function (Fig. 31.8). In addition to the response at τ_1 = 20 min following the solar wind arrival at the magnetopause and representing the direct driving of ionospheric DP2 currents by magnetospheric convection, in intermediate activity levels a second response appears at τ_2 = 60 min. The second response represents the substorm "unloading", or release of energy stored in the magnetotail. For strong driving the second peak disappears: there is not sufficient time for the energy to be stored and released later. The τ_1 and τ_2 time scales are averages; in actuality the growth phase prior to the second peak can be significantly longer, depending on how low the solar wind driving and energy storate rate is. Generally, the response is more complicated and nonlinear as we examine smaller ranges of activity level [Vassiliadis et al., 1995, 1996a].

The activity-dependent response shows that there is increased dissipation of the solar wind energy during substorms. Therefore the geomagnetic field has, at long time scales, a variation similar to that of the solar wind and IMF drivers, but for substorm time scales it is significantly modified.

The energy "loading-unloading" cycle during a substorm takes many forms in addition to the increased dissipation at ionospheric altitudes and intensification of regional currents such as the auroral electrojets. In the magnetotail, increased dissipation leads to severance and reconnection of magnetic field lines, during which a plasmoid, or magnetic island, is formed and is released to the solar wind (Fig. 31.9).

The nonlinear dynamics of the dissipation in the magnetotail, as well as the ionosphere, have given rise to the dripping faucet model of the magnetosphere [Baker et al., 1990] which is based on the classic nonlinear system [Shaw, 1984]. The dripping-faucet model led to more sophisticated models of the electrodynamics and thermodynamics of the global magnetosphere [Klimas et al., 1992; 1994; Horton and Doxas, 1996, 1998] (see also the chapter by Baker et al. in this book). In other approaches the geomagnetic variation during substorm cycle has been modeled as a phase transition, based on large-scale correlation of the auroral

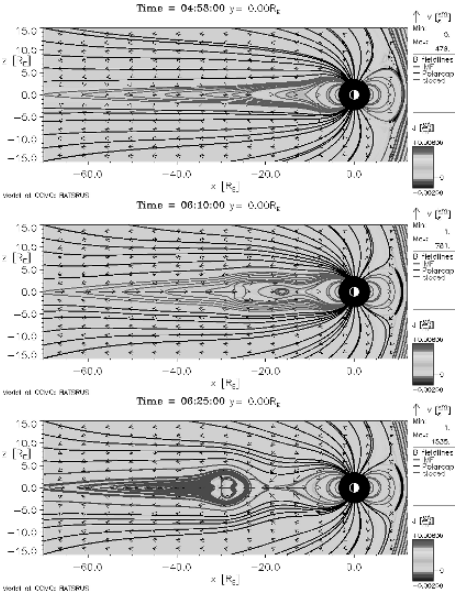


Fig. 31.9. Storage-release response of the magnetosphere during an isolated substorm simulated by the BATS-R-US code. Color contours represent the current density while short arrows represent the velocity field and lines the magnetic field. Note the formation and separation of the plasmoid from the initially extended magnetotail (after [Kuznetsova et al., 2006]). (A color version of this figure appears between pages 196 and 197).

Table 1. Modes of response of the magnetosphere.

Mode	Duration	Recurrence
Magnetic storm	2-7 days	1 month
Convection events	Hours-days	Random
Isolated substorm	2-4 hours	>4 hours
Sawtooth injections	1-2 hours	1-2 hrs for 10s of hours
Pseudobreakups	0.5-1 hours	-
LF waves/pulsations	1 sec-20 min	Random

electrojet indices and their time derivatives with the solar wind input, [Sitnov et al., 2000] and in terms of self-organized criticality, based on its scale-free distributions [Uritsky et al., 2001b].

The substorm is one of the many complex modes of response in the magnetosphere that determine the geomagnetic field variation. The presence of such spatio-

temporal modes of response changes the geomagnetic field "output" in a qualitative manner from the original solar wind input. In addition to the nonlinearity of the response, there are other effects that determine the field output such as the displacement of one or more current systems in geomagnetic coordinates. Nonlinear effects such as saturation and storage-release are common to magnetospheric and ionospheric regions and determine the changes from the solar wind input to the geomagnetic field "output". The presence of discrete response modes involves one or more storage-release effects. Other modes of response are shown in Table 31.1. Each one of these nonlinear modes contributes differently to the multifractal spectrum of the field and indices.

6 State Space Modeling

The dynamic, multifractal nature of the geomagnetic field is an indication of the modeling methods needed to represent it. The state space approach is one of these methods and as discussed elsewhere [Vassiliadis et al., 2000; Vassiliadis, 2006] it comprises quantitative and visualization methods, and can be combined with conventional, static models of the field.

We represent the field with local-linear (nonlinear) ARMA models of the type

$$B\left(t_k\right) = \sum_{i=1}^{m} a_i^{(LL)} B\left(t_{k-i}\right) + \sum_{j=0}^{l} b_j^{(LL)} I\left(t_{k-j}\right) \quad (7)$$

where B is a measure of geomagnetic activity such as the field, an index, or a time derivative of either one, and I(t) is a solar wind input, for instance $v_{sw}B_s$ or one of the aforementioned input variables. The state space for this model involves both geomagnetic and solar wind measurements (Fig. 31.10). After the model coefficients have been computed, the local-linear model (7) is used to nowcast and forecast the geomagnetic activity. Fig. 31.11 shows an example interval of very high prediction accuracy for the AL index.

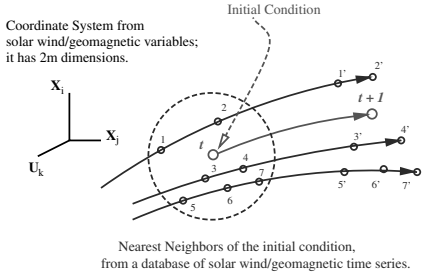


Fig. 31.10. State space prediction: the local-linear model is based on the nearest neighbors of the present magnetospheric and solar wind state.

Models such as (5) and (7) have also been applied to field and time-derivative modeling purposes [Vassiliadis et al., 2002]. Magnetic activity levels at nearby sites (where nearby is of the order of 100 km) have only small differences in amplitude or phase. We have modeled magnetic variations measured at the IMAGE magnetometer array (Fig. 31.12a) whose 27 magnetometers are deployed over a latitudinal range of 58-79° and where the magnetic field data are divided by magnetic local time (MLT) λ. In this way we have established a (λ,θ) grid of virtual magnetometers at whose vertices the geomagnetic North-South component B_x is represented by an FIR model similar to (5):

$$B_\theta(t;\lambda,\theta) = \int_0^\infty H^{B_\theta}_{v_{sw}B_s}(\tau;\lambda,\theta)\, I(t-\tau)\, d\tau \quad (8)$$

where I(t) is either a geomagnetic parameter such as the polar cap geomagnetic index [Vassiliadis et al., 1996b]; or an interplanetary input of the types discussed above. The B_y and B_z components are modeled similarly to (8).

We have also modeled the $\partial \mathbf{B}/\partial t$ variation in the same way. The following is an FIR model of the North-South component of the time derivative:

$$\frac{\partial B_\theta(t;\lambda,\theta)}{\partial t} = \int_0^\infty H^{\partial B_\theta/\partial t}_{v_{sw}B_s}(\tau;\lambda,\theta)\, I(t-\tau)\, d\tau \quad (9)$$

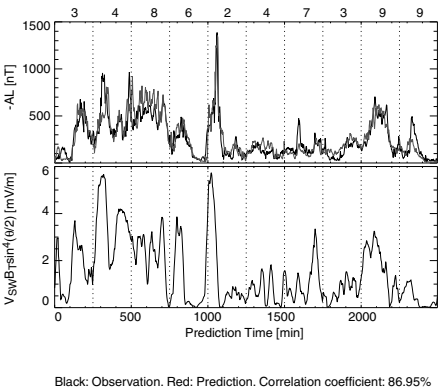


Fig. 31.11. Prediction of the AL index using a local-linear state space model. a. Observed (red) and model predicted (black) AL. b. The IEF $v_{sw}B_T \sin^4(\theta/2)$ used as a driver. (A color version of this figure appears between pages 196 and 197). (From [Vassiliadis et al., 2002]).

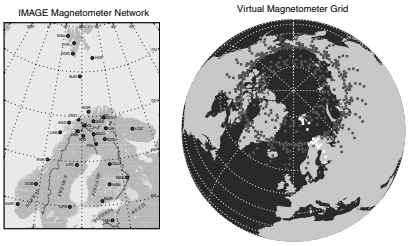


Fig. 31.12. a. Distribution of IMAGE array magnetometers in Fennoscandia. b. Virtual magnetometer grid developed for all MLT from the IMAGE array measurements.

Generally the input variables I(t) in the two models are different. For instance, the North-South component responds to the IEF $v_{sw}B_s$, but its time derivative is more strongly determined by the solar wind plasma velocity v_{sw} [Weigel et al., 2003].

The impulse response function for model (8) has several similarities to the impulse response for AL especially for nightside MLT. Shown in Fig. 31.13 is the response at midnight MLT driven by the polar cap index, I(t)=PC(t). Note that the response peaks are approximately $\Delta\tau$=40 minutes apart, similar to the $\Delta\tau$ separation between directly driven and loading-unloading modes in (6) of Fig. 31.8. The second peak τ_2 corresponds to the substorm phases of expansion (unloading) and recovery. Note also that the first peak of $H^{B_\theta}_{PC}$ occurs at τ_1=-2 min as opposed to τ_1=20 min for the $H^{AL}_{v_{sw}B_s}$ response: temporal shifts such as this are due to the time difference between PC and $v_{sw}B_s$ which is approximately 20 min.

The impulse response functions give information about the temporal as well as spatial development of the modes of response. In Fig. 31.13, we discern spatial features such as an equatorward expansion of the high impulse response for τ<0 following by a retreat to poleward latitudes for τ>0. In the spatiotemporal model built from these responses as functions of the geomagnetic coordinates we find that the peaks corre-

spond to the intensifications of currents (such as electrojets and field-aligned currents). The geomagnetic activity is strongly dependent on magnetic latitude. In the auroral zone (58-73°) we find a strong response field whereas in the polar cap and sub-auroral latitudes, to the zone's North and South, respectively, the response is weaker and more transient.

The variation of $\partial B_\theta(t;\lambda,\theta)/\partial t$ is fairly complex as well, and it too can be viewed as the response on several different interplanetary and/or geomagnetic inputs. The response to the PC index is characterized by a spectrum of oscillation as Fig. 31.14 shows for midnight MLT. This is a continuous spectrum of oscillations ranging from a few minutes to tens of minutes and hours. Signatures of current systems are consistent for both B_θ and $\partial B_\theta(t;\lambda,\theta)/\partial t$. Maps of activity are synthesized from the individual filters as well.

The above models of the magnetic field measures are verified against observations taken from several different intervals. Comparisons can be made at specific vertices of the virtual magnetometer grid, or in terms of instantaneous maps. Fig. 31.15 shows the prediction capability of model (8) at a moderately strong (B_θ=-683 nT) substorm that took place on October 1, 2003. We note that stations at far North or South latitudes are not represented accurately by the model while there is good agreement for most of the stations, including all of the auroral zone. The agreement is measured quantitatively using metrics such as the correlation coefficient and the prediction error.

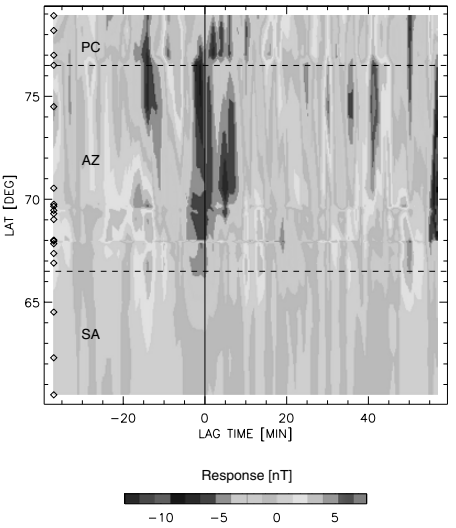


Fig. 31.13. The impulse response function for the B_θ (North-South) component of the geomagnetic field to the polar cap geomagnetic index at local midnight. PC: polar cap; AZ: auroral zone; SR: subauroral region. (A color version of this figure appears between pages 196 and 197).

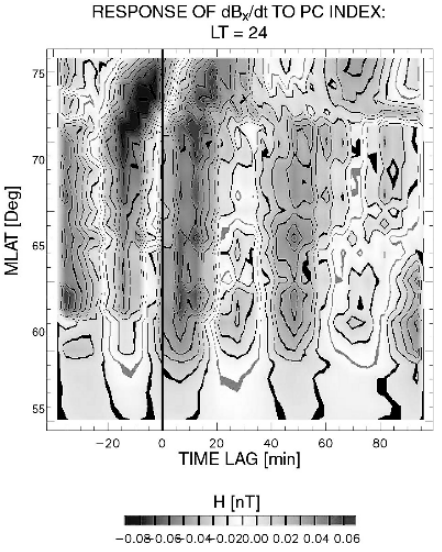


Fig. 31.14. As in Fig. 31.13, but for the time derivative $\partial B_\theta/\partial t$. (A color version of this figure appears between pages 196 and 197).

In summary we have seen how time variations in all 3 geomagnetic field measure types can be represented by linear and nonlinear deterministic models driven with interplanetary or magnetospheric activity inputs. Thus, the field can be seen as due to a set of nonlinear responses that organize the solar wind variability and can be represented by dynamical models.

7 Discussion and Summary: Modal or Multifractal?

We have seen that the geomagnetic variations, ranging from minutes to years, can be approximated by a multifractal representation. The multifractal description is also useful for derived measures such as magnetic indices and time derivatives.

For timescales longer than 3.5-5 hours, the multifractal spectra of geomagnetic variables are similar to spectra of certain types of interplanetary activity and therefore the solar wind-magnetosphere coupling can be considered linear or weakly nonlinear. However, at shorter timescales, storage-release geospace phenomena cannot be neglected and the response is strongly nonlinear as seen from different time series analyses [Tsurutani et al., 1990; Uritsky et al., 2001a; Pulkkinen et al., 2006].

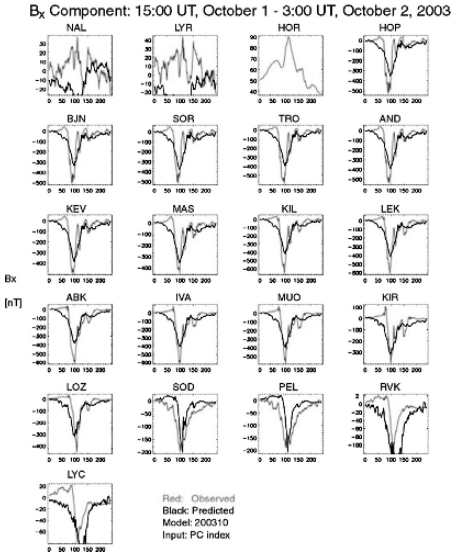


Fig. 31.15. Verification of model (8) for a magnetospheric substorm on October 1-2, 2003. Observed B_θ is shown in red, model-predicted in black. (A color version of this figure appears between pages 196 and 197). (From [Vassiliadis et al., 2002]).

Therefore it seems that the multifractal description has different interpretations at different scales. At short timescales, multifractal scalings is related to the stochastic variability of the magnetospheric field and plasma parameters, for instance in the magnetotail [Borovsky et al., 1997]. The invariance has been explained by turbulence and/or self-organized criticality.

On the other hand, a dynamical description can be directly coupled to the solar wind input, and can be used to represent the nonlinear response of magnetospheric and ionospheric plasmas. The response modes are important in modeling and understanding geomagnetic field variations and some of their statistical properties. Modes appear in a variety of spatial, temporal, and energy scales, and represent sets of plasma processes driven by multivariate rather

than univariate inputs [e.g., Vassiliadis et al., 2005]. However, the dynamic model description does not generally capture the fractal character of the field so a synthesis is necessary.

In summary, therefore, the time-complexity of the magnetic field is a balance between the combination of external (multiscale) driving plus a nonlinear deterministic response, and additional stochastic effects. In some cases the nonlinear response is the best way to explain the field variability: over the last two decades, nonlinear-dynamic models are capable of reproducing the time dependence of index and field variations, and are clear improvements over their linear counterparts [Vassiliadis et al., 1995, 1996a; Gleisner and Lundstedt, 1997; Valdivia et al., 1999a, b]. Predictability is high for indices such as AU, D_{st}, and K_p, and the polar and subauroral ground field. It is low to moderate for AL, AE, and the auroral-region ground field. In other cases, such as the field in the plasma sheet, a stochastic description is still the only one available at present.

Acknowledgments. The chapter is based upon work supported by the National Science Foundation under Grant No. 0602652.

References

Burlaga, L.F., Interplanetary Magnetohydrodynamics, Oxford University Press, 1995.

Baker, D.N., A.J. Klimas, R.L. McPherron, J. Buechner, The evolution from weak to strong geomagnetic activity: an interpretation in terms of deterministic chaos, Geophys. Res. Lett. 17, 1, 41-44, 1990.

Bargatze, L.F., D.N. Baker, R.L. McPherron, and E.W. Hones, Jr., Magnetospheric impulse response for many levels of geomagnetic activity, J. Geophys. Res. 90, A7, 6,387-6,394, 1985.

Borovsky, J.E., R.J. Nemzek, and R.D. Belian, The occurrence rate of magnetospheric-substorm onsets: random and periodic substorms, J. Geophys. Res. 98, A3, 3807-3814, 1993.

Consolini, G., M.F. Marcucci, and M. Candidi, Multifractal structure of auroral electrojet index data, Phys. Rev. Lett. 76, 4082-4085, 1996.

Clauer, C.R., The technique of linear prediction filters applied to studies of solar wind-magnetosphere coupling, in: Solar Wind-Magnetosphere Coupling, Y. Kamide and J.A. Slavin (eds.), 39-57, Terra Scientific, 1986.

Klimas, A.J., D.N. Baker, D.A. Roberts, D.H. Fairfield, J. Buchner, A nonlinear dynamics analogue model of geomagnetic activity, J. Geophys. Res. 97, A8, 12,253-12,266, 1992.

Klimas, A.J., D. N. Baker, D. Vassiliadis, and D. A. Roberts, Substorm recurrence during steady and variable solar wind driving: Evidence for an normal mode in the unloading dynamics of the magnetosphere, J. Geophys. Res. 99, A8, 14,855-14,861, 1994.

Mayaud, P.N., Derivation, meaning and use of geomagnetic indices, Geophys. Monogr. 22, American Geophysical Union, Washington, DC, 1980.

Menvielle, M., and A. Berthelier, The K-derived planetary indices: description and availability, Rev. Geophys. 29, 415, 1991.

Gleisner, H., and H. Lundstedt, Response of the auroral electrojets to the solar wind modeled with neural networks, J. Geophys. Res. A7, 102, 14269-14278, 1997.

Horton, W., and I. Doxas, A low-dimension energy-conserving state space model for substorm dynamics, J. Geophys. Res. 101, A12, 27,223-27,237, 1996.

Horton, W., and I. Doxas, A low-dimensional dynamical model for the solar wind driven geotail-ionosphere system, J. Geophys. Res. 103, A3, 4561-4572, 1998.

Kuznetsova, M.M., M. Hesse, L. Rastaetter, T. Gombosi, D. Zeeuw, G. Toth, Collisionless reconnection in global modeling of magnetospheric dynamics, Eos Trans. AGU, 87 (52), Fall Meet. Suppl., Abstract SM34B-03, 2006.

Lui, A.T.Y., A synthesis of magnetospheric substorm models, J. Geophys. Res. 96, A2, 1849-1856, 1991.

Osborne, A.R., and A. Provenzale, Finite correlation dimension for stochastic systems with power-law spectra, Physica D35, 3, 357-381, 1989.

Pavlos, G.P., M.A. Athanasiu, D. Kugiumtzis, N. Hatzigeorgiou, A.G. Rigas, and E.T. Sarris, Nonlinear analysis of magnetospheric data. Part 1. Geometric characteristics of the AE index time series and comparison with nonlinear surrogate data, Nonlin. Proc. Geophys. 6, 51-65, 1999.

Pavlos, G.P., D. Kugiumtzis, M.A. Athanasiu, N. Hatzigeorgiou, D. Diamantidis, and E.T. Sarris, Nonlinear analysis of magnetospheric data. Part II., Dynamical characteristics of the AE index time series and comparison with nonlinear surrogate data, Nonlin. Proc. Geophys. 6, 79-98, 1999.

Prichard, D., and C. Price, Spurious dimension estimates from time series of geomagnetic indices, Geophys. Res. Lett. 19, 15, 1623-1626, 1992.

Prichard, D., and C.P. Price, Is the AE index the result of nonlinear dynamics?, Geophys. Res. Lett. 20, 2817-2820, 1993.

Pulkkinen, A., A. Klimas, D. Vassiliadis, V. Uritsky, E. Tanskanen, Spatiotemporal scaling properties of the ground geomagnetic field variations, J. Geophys. Res. 111, A03305, doi:10.1029/2005JA011294, 2006a.

Pulkkinen, A., A. Klimas, D. Vassiliadis, and V. Uritsky, Role of stochastic fluctuations in the magnetosphere-ionosphere system: a stochastic model for the AE index variations, J. Geophys. Res. 111, A10218, doi: 10.1029/2006JA011661, 2006b.

Pulkkinen, T.I., D.N. Baker, L.A. Frank, J.B. Sigwarth, H.J. Opgenoorth, R. Greenwald, E. Friis-Christensen, T. Mukai, R. Nakamura, H. Singer, G.D. Reeves, M. Lester, Two substorm intensifications compared: onset, expansion, and global consequences, J. Geophys. Res. 103, A1, 15-28, 1998.

Roberts, D.A., D.N. Baker, A.J. Klimas, and L.F. Bargatze, Indications of low-dimensionality in magnetospheric dynamics, Geophys. Res. Lett. 18, 151, 1991.

Sitnov, M.I., A. S. Sharma, K. Papadopoulos, D. Vassiliadis, J. A. Valdivia, A. J. Klimas, and D. N. Baker, Phase transition-like behavior of the magnetosphere during substorms, J. Geophys. Res. 105, A6, 12955-12974, 2000.

Shan, L.-H., P. Hansen, C.K. Goertz, R.A. Smith, Chaotic appearance of the AE index, Geophys. Res. Lett. 18, 2, 147-150, 1991.

Shaw, R., The dripping faucet as a model chaotic system, Aerial Press, Santa Cruz, CA, 1984.

Sornette, D., Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools, Springer, 2004.

Takalo, Jouni, Ph.Lic. thesis, Report LT 3/93, Dept. Physics, Univ. of Jyvaskyla, 1993.

Tsurutani, B.T., Goldstein, B.E., M. Sugiura, T. Iyemori, W.D. Gonzalez, The nonlinear response of AE to the IMF Bs driver – A spectral break at 5 hours, Geophys. Res. Lett. 17, 279-282, 1990.

Uritsky, V.M., A. J. Klimas, D. Vassiliadis, Comparative study of dynamical critical scaling in the auroral electrojet index versus solar wind fluctuations, Geophys. Res. Lett. 28, 19, 3890-3812, 2001a.

Uritsky, V., M. Pudovkin, and A. Steen, Geomagnetic substorms as perturbed self-organized dynamics of the magnetosphere, J. Atmos. Sol.-Terr. Phys. 63, 1415-1424, 2001b.

Uritsky, V.M., A.J. Klimas, and D. Vassiliadis, Critical finite-size scaling of energy and lifetime probability distributions of auroral emissions, Geophys. Res. Lett. 33, LXXXXX, doi: 10.1029/2005GL025330, 2006.

Valdivia, J.A., D. Vassiliadis, A. J. Klimas, and A. S. Sharma, The spatio-temporal activity of magnetic storms, J. Geophys. Res. 104, A6, 12,239-12,250, 1999a.

Valdivia, J.A., D. Vassiliadis, A. J. Klimas, and A. S. Sharma, Modeling the spatial structure of the high-latitude magnetic perturbations and the related current systems, Phys. Plasmas 6, 11, 4185-4194, 1999b.

Vassiliadis, D., A. S. Sharma, T. E. Eastman, and K. Papadopoulos, Low-dimensional chaos in magnetospheric activity from AE time series, Geophys. Res. Lett. 17, 1841-1844, 1990.

Vassiliadis, D., A. J. Klimas, D. N. Baker, and D. A. Roberts, A description of solar wind-magnetosphere coupling based on nonlinear filters, J. Geophys. Res. 100, A3, 3495-3512, 1995.

Vassiliadis, D., A. J. Klimas, D. N. Baker, D. A. Roberts, The nonlinearity of models of the vBSouth - AL coupling, J. Geophys. Res., 101, A9, 19779-19787, 1996a.

Vassiliadis, D., V. Angelopoulos, D. N. Baker, and A. J. Klimas, The relation between northern polar cap and auroral electrojet geomagnetic indices in the wintertime, Geophys. Res. Lett. 23, 20, 2781-2784, 1996b.

Vassiliadis, D., System identification, modeling, and predictions for space weather environments, IEEE Trans. for Plasma Phys. 28, 6, 1944-1955, 2000.

Vassiliadis, D., A. J. Klimas, B.-H. Ahn, R. J. Parks, A. Viljanen, and K. Yumoto, High-latitude electrodynamics from a multi-array nonlinear geomagnetic model, in: Space Weather Study Using Multipoint Techniques, L.-H. Lyu (ed.), COSPAR Colloquia Series 12, Elsevier, Amsterdam, 2002.

Vassiliadis, D., S.F. Fung, and A.J. Klimas, Interplanetary and magnetospheric state parameters for the radiation belt electron flux, J. Geophys. Res. 110, A4, Art. No. A04201, 2005.

Vassiliadis, D., System theory for geospace plasma dynamics, Reviews of Geophysics, June 2006.

Viljanen, A., O. Amm, and R. Pirjola, Modeling geomagnetically induced currents during different ionospheric situations, J. Geophys. Res. 104, A12, 28,059-28,071, 1999.

Weigel, R.S., A.J. Klimas, D. Vassiliadis, Solar wind coupling to and predictability of ground magnetic fields and their time derivatives, J. Geophys. Res. 108, A7, 1298, 10.1029/2002JA009627, 2003.

Index