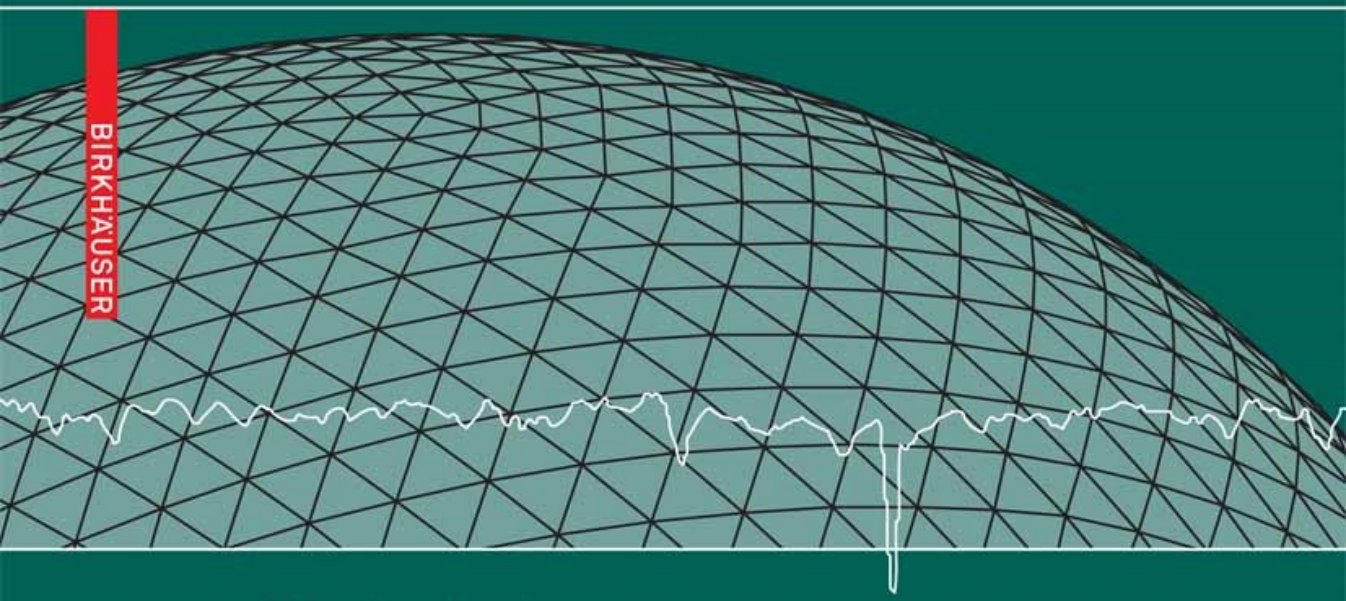
Earth Sciences and Mathematics

Volume I

Edited by
Antonio G. Camacho
Jesús I. Díaz
José Fernández



pageoph topical volumes

BIRKHÄUSER

Earth Sciences and **Mathematics**

Volume I

Edited by
Antonio G. Camacho
Jesús I. Díaz
José Fernández

Birkhäuser
Basel · Boston · Berlin

Reprint from Pure and Applied Geophysics
(PAGEOPH), Volume 165 (2008) No. 6

Editors:
Antonio G. Camacho
José Fernández
Instituto de Astronomía y Geodesia
(CSIC-UCM)
Facultad Ciencias Matemáticas
Universidad Complutense Madrid
Ciudad Universitaria
Plaza de Ciencias 3
28040 Madrid
Spain
Email: antonio_camacho@mat.ucm.es
jose_fernandez@mat.ucm.es

Jesús I. Díaz
Instituto de Matemática Interdisciplinar (IMI)
Departamento de Matemática Aplicada
Facultad Ciencias Matemáticas
Universidad Complutense Madrid
Ciudad Universitaria
Plaza de Ciencias 3
28040 Madrid
Spain
Email: diaz.racefyn@insde.es

Library of Congress Control Number: 2008932710

Bibliographic information published by Die Deutsche Bibliothek:
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>

ISBN 978-3-7643-8906-2 Birkhäuser Verlag AG, Basel · Boston · Berlin

ISBN 978-3-7643-8906-2 e-ISBN 978-3-7643-8907-9

9 8 7 6 5 4 3 2 1 www.birkhauser.ch

Contents

Pure appl. geophys. 165 (2008) 997–1001
0033–4553/08/060997–5
DOI 10.1007/s00024-008-0343-7

Pure and Applied Geophysics

Introduction: Linking Earth Sciences and Mathematics

A. G. Camacho,[1] J. I. Díaz,[2] and J. Fernández[1]

Knowledge of the Earth's structure and dynamics calls for a multi-disciplinary study that makes use of the most advanced methods of Physics, Chemistry, Mathematics and Information Technology, in the framework, or in a close collaboration with, the different branches of Earth Sciences such as Geology, Geophysics and Geodesy. The research to be developed includes subjects ranging from data acquisition, both with traditional techniques and with the most advanced resources of our time; data treatment and processing; to the development of new modelling methodologies for the simulation and reproduction and prediction of the terrestrial processes on a local, regional and, by far the most ambitious, global scale.

The large amount of (geological, geophysical and geodetic) high precision observation data about the Earth available acquired both from the planet itself and from space, grows increasingly. Currently one can obtain huge amounts of high precision data that cover the widest areas desired. Often the time or space distribution of these data are almost continuous, and considerable of data has been obtained by unconventional techniques. In view of the privileged situation at present, we must reconsider the connection and utilization of that abundant data with the more theoretical studies that offer new and more refined mathematical models, capable of making the most of the technological breakthroughs in observation. In the words of Jacques-Louis Lions (1928–2001): "*If we accept that any mathematical models isolated of any experimental data have no predictive value, similarly, we must recognize that the most abundant data banks without good mathematical models produce nothing more than confusion*".

What is required is to develop new mathematical, analytical and numerical, models and methods for data processing and interpretation, considering their ever-increasing quality, variety in origin (terrestrial and space), type (it is becoming more and more possible to measure larger numbers of parameters simultaneously that can be related to

[1] Instituto de Astronomía y Geodesia (CSIC-UCM), Fac. C. Matemáticas, Ciudad Universitaria, Plaza de Ciencias, 3, 28040 Madrid, Spain. E-mail: antonio_camacho@mat.ucm.es, jose_fernandez@mat.ucm.es

[2] Instituto de Matemática Interdisciplinar (IMI) and Departamento de Matemática Aplicada, Fac. C. Matemáticas, Universidad Complutense de Madrid, Ciudad Universitaria, Plaza de Ciencias, 3, 28040 Madrid, Spain.

one another), time (data acquired sporadically or continuously) and space extension (going from disperse spots to almost continuous observation in space).

The generalized use of the data obtained in observing the Earth from space (ESA, 2008, NASA, 2008), the ever-increasing number of problems in which they are applicable, and the need to combine them with the data acquired on Earth, poses new problems both in the field of the statistical processing of the data and in their use by society. All of this requires the involvement of specialists in such differing subjects as decision-making support, operational research and artificial intelligence.

Furthermore, the efficient use of all this terrestrial and space data also requires the help of the most sophisticated mathematical models that permit the correct interpretation of these data together. At present, numerous problems remain to be solved in this field, and there is huge demand for inversion models and techniques, and for other mathematical tools, among the community of Earth Sciences specialists. One very illustrative example of that interaction with sophisticated mathematical techniques refers to the application of the latest mathematical developments on complex systems and chaos theory to the specific case of Earth Sciences. Inversely, Geosciences gives to mathematicians new and difficult problems which some times need new developments to be at least partially solved. Therefore a closer cooperation between researchers from both fields became day by day more necessary and is a basic aspect integral to carrying out many works devoted to understand and solve problems related with geodynamics, natural hazards, global change, etc.

From the approach of the need for a greater application and integration of mathematics in the study of the Earth, the name of Geomathematics has been put forward to combine the research and works that seek to develop and incorporate new methods, approaches and solutions from different areas of mathematics, such as Statistics, Operational Research, Artificial Intelligence and more in general from Applied Mathematics, with special emphasis on its Modelling, Analysis, Numerical and Computational Approximation methods and processes, and its Control Theory techniques.

Yet that mathematical study of the Earth could never be carried out successfully without the close collaboration with the specialists of all branches (Geology, Geophysics and Geodesy) of Earth Sciences. The international (not to mention planetary) dimension at which these contacts must be maintained is evident (see e.g., IUGS, 2008; IUGG, 2008). This philosophy, many aspects of which are already in place, will make it possible to tackle the most current and ambitious scientific challenges that our society requires and will also trigger important advances in the frontier of knowledge and culture that will redound positively to the welfare of society.

Geomathematical research involves such diverse studies as flow models (porous means, glaciers, etc.), sedimentation and diagenesis, global change models, wave propagation, classification of the Earth's surface, riskmap analysis, parameter sensitivity analysis in inverse problems, stochastic models for processing terrestrial and space data, direct deterministic models, chaos theory applied to Earth Sciences, geostatistical

software, poor, incomplete or truncated information problems, time series analysis, information dimension reduction, information representation, interfaces for reports in specific fields, handling linguistic information, nonlinear processes in Earth Sciences, study of geological structures and phenomena with invariance of scale by means of self-organized criticality methods and fractal-type space distributions, etc.

Considering this perspective of the evolution of Earth Sciences and with the idea of fostering the aforementioned collaboration, the Complutense International Seminar on "Earth Sciences and Mathematics" was organized and held in Madrid at the Faculty of Mathematical Sciences of the Universidad Complutense de Madrid from 13–15 September 2006. Scientists from both fields, Mathematics and Earth Sciences, took part in this International Seminar, addressing scientific problems related with our planet from clearly complementary approaches, seeking to gain and learn from this dual approach and proposing closer collaboration in the near future.

This volume is the first Topical Issue on "Earth Sciences and Mathematics" and contains 11 papers, most of which were presented at the International Seminar. They address different topics as are deformation modelling applied to natural hazards, inverse gravimetric problem to determine 3D density structure, advanced differential SAR interferometry, climate change, geomagnetic field, Earthquake statistics, meteorological studies using satellite images, climate energy balance models, study of soils properties, multifractal data sets, etc.

HOLLIDAY *et al*. present a well-written and concise review of the literature and some of the recent work by the group about the probabilistic risk of earthquake occurrence. The paper introduces a very interesting stochastic model for seismicity, alternative to the popular ETAS model: The branching aftershock sequence (BASS) model. It is accompanied by tutorial-type examples.

The paper by BERMEJO *et al*. introduces a numerical algorithm for the study of some climate energy balance models which consists of a two-dimensional nonlinear parabolic problem on the 2-sphere with the albedo terms formulated according to Budyko as a bounded maximal monotone graph in R^2. The numerical model combines the first order Euler implicit time discretization scheme with linear finite elements for space discretization given by quasi-uniform spherical triangles.

EFF-DARWICH *et al*. outline an interesting data supply option: The employment of tiltmeter records from a solar telescope (THEMIS) in Tenerife (Canary Islands) as a complementary technique to the geodetic system for continuous monitoring of ground deformation in this volcanic island. Authors describe the wavelet procedure used for the data analysis, and show significant signal tilts, which are interpreted as associated with a major submarine fault.

The paper by ÁLVAREZ *et al*. proposes a mathematical model (of a variational nature and leading then to some nonlinear Partial Differential Equations) to analyze sequences of a two-dimensional multichannel meteorological satellite image including visible, temperature and water vapor channels of great relevance in the study of cloud structures and their displacements.

The paper by BERRINO and CAMACHO presents a structural study of the Mt. Vesuvius volcano, its magmatic system and the entire Neapolitan area by applying a new 3D method for gravity inversion to the available Bouguer map. The method describes the subsurface structures in terms of several sub-horizontal layers, each representing a specific geological formation.

GARCÍA-GUTIERREZ and MARTÍN develop the log-self-similar model which may be a useful tool to simulate particle size distribution in solids by using some fractal techniques which can be useful for the construction of pedotransfer functions related to other soil properties when textural information is limited to modest textural data.

RODELSPERGER *et al.* examine the impact of steric sea-level change on the gravity field in the next 2000 years according to two scenarios of CO_2 emissions. As expected, the authors find that this impact is negligible compared to mass changes in the ocean. Nonetheless this study seems to be the first one providing numerical evidence of that result, and the methodology presented here could be modified to be applied to other studies.

The paper by MARTÍN and REYES deals with the interaction of irregular, winding, dragging paths through soil complex distributions. A mathematical modelling of the interplay between multifractal distributions of mineral/pollutants in soil and fractal pore networks is presented. A Hölder path is used as a model of soil pore network and a multifractal measure as a model of soil complex distribution. They show that the Hölder exponent of the path and the entropy dimension of the distribution may be used to quantify such interplay.

BLANCO *et al.* present an overview of the advanced differential SAR interferometry (DInSAR) technique referred to as Coherent Pixel Technique (CPT). Besides an interesting improvement (multi-layer processing), the authors present several valuable deformation results of wide areas using the differential SAR technique, showing that this tool has major potential to be used for detecting and monitoring deformation phenomena, as well as to better understand the geological mechanisms that provoke them.

The paper by BOGACHEV *et al.* analyzes the statistics of return intervals to better understand the occurrence of extreme events. The major issue is to determine some general "scaling" relations between the return intervals at low and high thresholds, which then allow extrapolation of the results to very large, extreme thresholds. They review former results for long-term correlated data sets and their most recent findings for multifractal data sets.

PAVÓN *et al.* extend an interesting and potentially very useful application of geomagnetic field modeling to archaeomagnetic dating, through the introduction of intensity data. The resulting regional archaeomagnetic model SCHA.DI.00-F, valid for the last 2000 years in the European region, is considered a valuable contribution to current discussions regarding the relationship between changes in the geomagnetic field and climate parameters.

A second volume will be published later in 2008 with more selected and reviewed papers presented at the Complutense International Seminar on "Earth Sciences and Mathematics".

We appreciate the splendid and generous work carried out by the many referees. They have worked in most of the cases, in the difficult intersection of two different fields as Earth Sciences and Mathematics. The reviewers have been: J. Almendros, M. Badii, M. Bebbington, A. Beliaev, P. Berardino, A. Bru, D. Chambers, D. F. Cook, A. Corral, A. Correig, D. Dong, B. Enescu, D. García, G. Gagneux, F. Giraldo, J. Gottsmann, G. Hetzer, R. C. A. Hindmarsh, A. Hooper, G. Houseman, T. Jahr, G. Jentzsch, E. Kerre, L. Kuchment, M. Laba, R. Lanari, J. Langbein, A. Lodge, A. Lombard, I. Main, V. C. Manea, S. McNutt, T. Mikumo, P. Moczo, M. Pacella, Y. Pachepsky, J. W. Parker, A. Peresan, P. Prats, J. M. Rey Simó, T. Sagiya, U. Schlink, C. Schoof, D. Seber, R. Shcherbakov, E. Sturkell, D. Salstein, K. F. Tiampo, D. Tarling, C. Vázquez Cendón, G. Wadge, C. Wicks, G. Zoeller, and D. Zupanski.

We would like to take this opportunity to thank the Vicerrectorate for International Relations of the Universidad Complutense de Madrid, the Spanish Council for Scientific Research, the Interdisciplinary Mathematical Institute (IMI), and the Royal Spanish Academy of Sciences for their support in organizing the International Seminar. This International Complutense Seminar was also partially supported with funds from research projects CGL-2004-21019-E and CGL2005-05500-C02-01. The editors also thank Renata Dmowska for the help, suggestions and support received during the editing process of this topical issue. And finally we wish to thank all the authors of this volume for their contributions.

References

ESA, European Space Agency- Observing the Earth (2008). http://www.esa.int/esaEO/index.html.
IUGG, International Union of Geodesy and Geophysics (2008). http://www.iugg.org/.
IUGS, International Union of Geological Sciences (2008). http://www.iugs.org/.
NASA, National Aeronautics and Space Administration (2008). http://www.nasa.gov.

Pure appl. geophys. 165 (2008) 1003–1024
0033–4553/08/061003–22
DOI 10.1007/s00024-008-0344-6

Pure and Applied Geophysics

A Review of Earthquake Statistics: Fault and Seismicity-Based Models, ETAS and BASS

James R. Holliday,[1,2] Donald L. Turcotte,[3] and John B. Rundle[1,2,3]

Abstract—There are two fundamentally different approaches to assessing the probabilistic risk of earthquake occurrence. The first is fault based. The statistical occurrence of earthquakes is determined for mapped faults. The applicable models are renewal models in that a tectonic loading of faults is included. The second approach is seismicity based. The risk of future earthquakes is based on the past seismicity in the region. These are also known as cluster models. An example of a cluster model is the epidemic type aftershock sequence (ETAS) model. In this paper we discuss an alternative branching aftershock sequence (BASS) model. In the BASS model an initial, or seed, earthquake is specified. The subsequent earthquakes are obtained from statistical distributions of magnitude, time, and location. The magnitude scaling is based on a combination of the Gutenberg-Richter scaling relation and the modified Båth's law for the scaling relation of aftershock magnitudes relative to the magnitude of the main earthquake. Omori's law specifies the distribution of earthquake times, and a modified form of Omori's law specifies the distribution of earthquake locations. Unlike the ETAS model, the BASS model is fully self-similar, and is not sensitive to the low magnitude cutoff.

1. Introduction

Deformation of the Earth's crust is responsible for the generation of earthquakes over a wide range of scales. In terms of the resulting seismicity, the Earth's crust is clearly a self-organizing complex system (Main, 1996; Rundle *et al.*, 2003). Despite this complexity, seismicity satisfies a number of universal scaling laws. These scaling laws have important implications for probabilistic seismic hazard analysis and earthquake forecasting. They also form the basis for a variety of models and simulations of earthquake activity. Examples include the epidemic type aftershock sequence (ETAS) model and the branching aftershock sequence (BASS) model. A comparison of these models will be a major focus of this paper.

Earthquakes constitute a major hazard on a worldwide basis. Although the locations of large earthquakes are concentrated near plate boundaries, they can occur within plate interiors. A specific example is the three large (magnitude ~ 7.7) earthquakes that occurred near New Madrid, Missouri in 1810 and 1811. A number of very large cities are

[1] Center for Computational Science and Engineering, University of California, Davis.
[2] Department of Physics, University of California, Davis.
[3] Department of Geology, University of California, Davis.

located very close to plate boundaries. Examples include Tokyo, Los Angeles, San Francisco, Seattle, Lima, Jakarta, and Santiago. Much of China is a diffuse plate boundary, and major earthquakes have caused large losses of life throughout this region. A recent example was in the 1976 Tangshan earthquake with some 500 000 deaths.

A major goal of earthquake research is to quantify the risk of occurrence of an earthquake of a specified magnitude, in a specified area, and in a specified time window. This is done and results in hazard maps. Historic and paleoseismicity are major constraints on seismic hazard assessments. Slip rates and recurrence intervals of earthquakes on recognized faults are specified in so far as data are available. Examples are the sequence of studies carried out by the working groups on California earthquake probabilities (FIELD, 2007b). These reports have formed the basis for establishing rates of earthquake insurance in California.

A second major goal of earthquake research is to specifically forecast or predict earthquakes. Many attempts have been made, but with only marginal success. A number of published forecasting algorithms involve the use of past seismicity. The occurrence of recent smaller earthquakes is extrapolated to forecast the occurrence of future larger earthquakes. We first consider the relative roles of fault-based models and seismicity-based models. These alternatives will be discussed in Sections 2 and 3.

A specific type of seismicity-based forecast models is the ETAS model. This model is discussed in Section 4. In Section 5 we introduce the BASS model, and in Section 6 we compare BASS to ETAS. We conclude that BASS is preferable because it is fully scale-invariant and satisfies the major accepted scaling laws of seismicity. In Section 7 we introduce a deterministic version of the BASS model and show that it exhibits Tokunaga scale-invariant, side-branching statistics. These statistics are also satisfied by river networks, diffusion-limited aggregation (DLA) clusters, branching in biology, and cluster growth in site percolation. In Section 8 we present a numerical simulation of an aftershock sequence using the BASS model. Finally, we state our conclusions in Section 9 and discuss the future of seismic hazard assessment.

2. *Fault-Based Models*

Fault-based models consider the earthquakes that occur on recognized active faults. These models are also known as renewal models. Renewal models require that the stress on an individual fault is "renewed" by the tectonic drive of plate tectonics. The simplest renewal model would be that of a single planar strike-slip fault subjected to a uniform rate of strain accumulation (plate motion). In this case, "characteristic" earthquakes would occur periodically. Clearly the Earth's crust is much more complex with faults present at all scales and orientations. This complexity leads to chaotic behavior and statistical variability.

An important question is whether the concept of quasi-periodic characteristic earthquakes is applicable to tectonically active areas. There is extensive evidence that

"characteristic" earthquakes do occur quasi-periodically on major faults. Many studies have been carried out to quantify the recurrence time statistics of these characteristic earthquakes (UTSU, 1984; OGATA, 1999; RIKITAKE, 1982). Recurrence time statistics can be characterized by a mean value, μ, and a coefficient of variation, C_v. The coefficient of variation is the ratio of the standard deviation to the mean. We have $C_v = 0$ for periodic characteristic earthquakes and $C_v = 1$ for a random distribution of recurrence times. ELLSWORTH *et al*. (1999) reviewed many examples of recurrence time statistics and concluded that $C_v \approx 0.5$ for characteristic earthquakes. Many probability distribution functions have been proposed for recurrence times, including the Weibull, lognormal, Brownian passage time, and gamma distributions.

Two major renewal simulation models have been developed. The first is "Virtual California" (RUNDLE *et al*., 2004, 2005, 2006). This is a geometrically realistic numerical simulation of earthquake occurring on the San Andreas fault system and includes all major strike-slip faults in California. The second model is the "Standard Physical Earth Model" (SPEM) developed by WARD (1992) and applied to characteristic earthquakes associated with subduction at the Middle American trench. This model was further developed and applied to the entire San Andreas fault system by GOES and WARD (1994), to the San Andreas system in southern California by WARD (1996), and to the San Andreas system in northern California by WARD (2000).

Both simulation models utilize backslip, with the accumulation of a slip deficit on each fault segment prescribed using available data. Ideally the tectonic drive would be applied directly to the edges of a region, say 200 km on each side of the San Andreas fault system. But the long-term evolution of this approach requires that faults become longer as slip accumulates. The resulting geometrical incompatibility leads to serious numerical problems. In the backslip models, continuous displacements are applied to all faults until the frictional constraints result in backslip (an earthquake). Both models "tune" the prescribed static friction to give recurrence times that are consistent with available data. In both models fault segments are treated as dislocations where characteristic earthquakes occur, and all fault segments interact with each other elastically utilizing dislocation theory. These chaotic interactions result in statistical distributions of recurrence times on each fault. The resulting coefficients of variation are measures of this interaction.

YAKOVLEV *et al*. (2006) utilized the Virtual California model to test alternative distributions of recurrence times. They concluded that the Weibull distribution is preferable and based its use on its scale invariance. The hazard rate is the probability distribution function (pdf) that a characteristic earthquake will occur at a time t_0 after the last characteristic earthquake. The Weibull distribution is the only distribution that has a power-law (scale-invariant) hazard function. YAKOVLEV *et al*. (2006) found that the coefficient of variation of the recurrence times of 4606 simulated great earthquakes on the northern San Andreas fault is $C_v = 0.528$. GOES and WARD (1994) using the SPEM simulator found that $C_v = 0.50 - 0.55$ on this fault. The two simulations are quite different, so the statistical variability appears to be a robust feature of characteristic

earthquakes. A similar simulation model for New Zealand has been given by ROBINSON and BENITES (1995, 1996.)

Renewal models have also formed the basis for three formal assessments of future earthquake probabilities in California. These assessments were carried out by the United States Geological Survey (WORKING GROUP ON CALIFORNIA EARTHQUAKE PROBABILITIES, 1988, 1990, 1995, 2003). A major problem with renewal models is that large earthquakes often occur on faults that were not previously recognized. Recent examples in California include the 1952 Kern County earthquake, the 1971 San Fernando Valley earthquake, the 1992 Landers earthquake, the 1994 Northridge earthquake, and the 1999 Hector Mine earthquake. At the times when these earthquakes occurred, the associated faults were either not mapped or were considered too small to have such large earthquakes. To compensate for this problem, renewal models often include a random level of background seismicity unrelated to recognized faults.

3. Seismicity-Based Models

An alternative approach to probabilistic seismic hazard assessment and earthquake forecasting is to use observed seismicity. The universal applicability of Gutenberg-Richter frequency-magnitude scaling allows the rate of occurrence of small earthquakes to be extrapolated to estimate the rate of occurrence and location of large earthquakes. This type of extrapolation played an important role in creating the national seismic hazard map for the United States (FRANKEL *et al.*, 1996).

A more formalistic application of this extrapolation methodology is known as a relative intensity (RI) forecast. This type of forecast was made on a worldwide basis by KOSSOBOKOV *et al.* (2000) and for California by HOLLIDAY *et al.* (2005). A related forecasting methodology is the pattern informatics (PI) method (RUNDLE *et al.*, 2002; TIAMPO *et al.*, 2002a, b; HOLLIDAY *et al.*, 2006a, b, 2007). This method was used by RUNDLE *et al.* (2002) to forecast $m = 5$ and larger earthquakes in California for the time period 2000–2010. This forecast successfully predicted the locations of 16 of the 18 large earthquakes that have subsequently occurred.

KEILIS-BOROK (1990, 2002) and colleagues utilized patterns of seismicity to make formal intermediate term earthquake predictions. Two notable successes were the 1988 Armenian earthquake and the 1989 Loma Prieta, California, earthquake. However, a number of large earthquakes were not predicted and the approach remains controversial. More recently, this group has used chains of premonitory earthquakes to make intermediate term predictions (SHEBALIN *et al.*, 2004; KEILIS-BOROK *et al.*, 2004). Again, moderate success was achieved.

It has also been proposed that there is an increase in the number of intermediate-sized earthquakes prior to a large earthquake. This proposal has been quantified in terms of an accelerated moment release (AMR) prior to a large earthquake. This approach has shown considerable success, retrospectively (BUFE and VARNES, 1993; BOWMAN *et al.*, 1998;

SAMMIS *et al*., 2004) but has not evolved into a successful prediction algorithm as of yet (GROSS and RUNDLE, 1998; MAIN, 1999).

Seismicity-based models are often referred to as clustering models. That is, clusters of small earthquakes indicate the future occurrence of larger earthquakes. The RI, PI, and AMR models clearly belong to this class. A rational for the application of clustering models is that the clustering is related to families of foreshocks, main shocks, and aftershocks. This rational forms the basis for the use of both the ETAS and the BASS models. It should be emphasized that neither model introduces renewal, the tectonic drive responsible for the energy input dissipated in earthquakes. One way to overcome this problem has been to introduce a random background seismicity to excite aftershocks. An alternative approach is to couple a clustering model, such as ETAS or BASS, to a fault-based model, such as Virtual California or SPEM.

4. ETAS

A clustering model that has been widely studied is the ETAS model. This approach was first formulated by KAGAN and KNOPOFF (1981). It is a statistical model based on applicable scaling laws. This model was further developed by OGATA and colleagues (OGATA, 1988, 1989, 1992, 1998, 1999, 2001a, b, 2004; OGATA *et al*., 1993, 2003; GUO and OGATA, 1997; OGATA and ZHUANG, 2006; ZHUANG and OGATA, 2006; ZHUANG *et al*., 2002, 2004; VERE-JONES, 2005). Modified versions of ETAS were introduced by HELMSTETTER, SORNETTE, and colleagues (HELMSTETTER, 2003; HELMSTETTER *et al*., 2003a, b, 2004, 2006; HELMSTETTER and SORNETTE, 2002a, b, 2003a, b, c, d; SAICHEV *et al*., 2005; SAICHEV and SORNETTE, 2004, 2005a, b, 2006a, b, c, 2007a, b; SORNETTE and HELMSTETTER, 2002; SORNETTE and WERNER, 2005a, b) and by LEPIELLO *et al*. (2007). Related models have been developed by Felzer and colleagues (FELZER *et al*., 2002, 2003, 2004), by Console and colleagues (CONSOLE and MURRU, 2001; CONSOLE *et al*., 2003, 2006), and by Gerstenberger and colleagues (GERSTERBERGER *et al*., 2004, 2005). Before discussing the details of the ETAS model, we will first introduce the BASS model.

5. BASS

An alternative to the ETAS model is the BASS model (TURCOTTE *et al*., 2007). As in the ETAS model, the BASS model recognizes that each earthquake has an associated sequence of aftershocks. Each main shock produces a sequence of primary aftershocks. Each of these aftershocks, in turn, produces second-order aftershocks. Each second-order aftershock can produce third-order aftershocks, and so forth. Statistically, a primary aftershock can be larger than the initial main shock. In this case the initial main shock becomes a foreshock, and the larger primary aftershock becomes the main shock of the

sequence. In principal, a higher-order aftershock could be the main shock. The probability of this occurring, however, is extremely small.

It has been demonstrated by many authors that the frequency-magnitude distribution of aftershocks satisfy the Gutenberg-Richter (GR) relation to a good approximation (GUTENBERG and RICHTER, 1954; SHCHERBAKOV *et al*., 2005). In the BASS formulation, we require that the frequency-magnitude distribution of each order of aftershocks satisfies the GR relation in the form

$$\log_{10}[N_d(\geq m_d)] = a_d - b_d m_d, \tag{1}$$

where m_d is the magnitude of a daughter earthquake, $N_d(\geq m_d)$ is the number of daughter earthquakes with magnitudes greater than or equal to m_d, and a_d and b_d are the a- and b-values of the distribution, respectively. Note that the b-value b_d for each sequence of aftershocks is not necessarily equal to the b-value for all aftershocks. This is due to the superposition of many generations of aftershock sequences for each parent earthquake and will be discussed in detail later in the paper.

In order to fully specify the frequency-magnitude distribution of a family of aftershocks, we apply the modified form of Båth's law (SHCHERBAKOV and TURCOTTE, 2004). As shown by SHCHERBAKOV *et al*. (2005), this formulation is closely related to that given by REASENBERG and JONES (1989), YAMANAKA and SHIMAZAKI (1990), and FELZER *et al*. (2002).

In its original form, Båth's law (BÅTH, 1965; VERE-JONES, 1969) states that the magnitude difference between a main shock and its largest aftershock Δm is nearly constant with a value near 1.2. SHCHERBAKOV and TURCOTTE (2004) introduced a new way of defining this difference and obtained a value Δm^* based on the entire distribution of aftershocks, not just the largest aftershock. It is required that the magnitude of the largest aftershock inferred from the GR relation is a fixed value Δm^* less than the magnitude of the parent earthquake, m_p:

$$N_d(\geq (m_p - \Delta m^*)) = 1. \tag{2}$$

With this condition we require (using Eq. (1) that $a_d = b_d (m_p - \Delta m^*)$ so that

$$\log_{10}[N_d(\geq m_d)] = b_d(m_p - \Delta m^* - m_d). \tag{3}$$

This relation fully specifies the frequency-magnitude distribution of each family of aftershocks (daughter earthquakes). However, this distribution implies an infinite number of small earthquakes. To eliminate this singularity, it is necessary to prescribe a minimum magnitude earthquake $m_{\min}$ that is to be considered.

We obtain this total number of aftershocks in a family by setting $m_d = m_{\min}$ in Eq. (3):

$$N_{dT} = N(\geq m_{\min}) = 10^{b_d(m_p - \Delta m^* - m_{\min})}. \tag{4}$$

This relation is the essential feature of the BASS model. Utilizing Eqs. (3) and (4), we obtain the cumulative distribution function P_{Cm} for the magnitudes of the daughter earthquakes:

$$P_{Cm} = \frac{N_d(\geq m_d)}{N_{dT}} = 10^{-b_d(m_d - m_{\min})}. \tag{5}$$

The magnitude m_d of each daughter earthquake is selected from this distribution. For each daughter earthquake a random value for P_{Cm} in the range $0 < P_{Cm} < 1$ is generated, and the magnitude of the earthquake is determined from Eq. (5). Note that there is a finite probability that a daughter earthquake can be larger than the parent earthquake. The probability that this occurs is obtained by substituting Eq. (4) into Eq. (5) and setting $m_d = m_p$ with the result

$$P_{Cm}(m_d > m_p) = 10^{-b_d \Delta m^*}. \tag{6}$$

Taking $b_d = 1$ and $\Delta m^* = 1$, we have $P_{Cm}(m_d > m_p) = 10\%$. With these values, a well defined foreshock would be expected about 10% of the time. This is in reasonable agreement with observed values of $13 \pm 5\%$ (REASENBERG, 1999). It should be emphasized that this value is independent of the choice for $m_{\min}$. The selection of a $m_{\min}$ is necessary, but the choice does not significantly influence the distribution of magnitudes above this cutoff.

Having specified the magnitude m_d of each daughter earthquake by selecting a random number P_{Cm} in the range $0 < P_{Cm} < 1$ and using Eq. (5) in the form

$$m_d = -\frac{1}{b_d} \log P_{Cm} + m_{\min}, \tag{7}$$

we next specify the time of occurrence of the daughter earthquakes.

We require that the time delay t_d until each daughter earthquake after the parent earthquake satisfies a general form of Omori's law (SHCHERBAKOV *et al.*, 2004):

$$R(t_d) = \frac{\mathrm{dN}_d}{\mathrm{d}t} = \frac{1}{\tau(1 + t_d/c)^p}, \tag{8}$$

where $R(t_d)$ is the rate of aftershock occurrence and τ, c, and p are parameters. The number of daughter aftershocks that occur after a time t_d is then given by

$$N_d(\geq t_d) = \int_{t_d}^{\infty} \frac{\mathrm{dN}_d}{\mathrm{d}t} = \frac{c}{\tau(p-1)(1 + t_d/c)^{p-1}}. \tag{9}$$

The total number of daughter earthquakes is obtained by setting $t_d = 0$ in Eq. (9) with the result

$$N_{dT} = \frac{c}{\tau(p-1)}. \tag{10}$$

It should be emphasized that this result is only valid for $p > 1$. If $p \leq 1$, the integral is not convergent and a maximum time must be specified. From Eqs. (9) and (10) we obtain the cumulative distribution function P_{Ct} for the times of occurrence of the daughter earthquakes:

$$P_{Ct} = \frac{N_d(\geq t_d)}{N_{dT}} = \frac{1}{(1 + t_d/c)^{p-1}}. \tag{11}$$

The time of occurrence t_d of each daughter earthquake is selected from this distribution. For each daughter earthquake a random value for P_{Ct} in the range $0 < P_{Ct} < 1$ is generated, and the time of occurrence of the earthquake is determined using Eq. (11) in the form

$$t_d = c(P_{Ct}^{-1/(p-1)} - 1). \tag{12}$$

The distribution of times is dependent only on the fitting parameters c and p. In this paper, these parameters for each generation of aftershocks are assumed to be equal. The values for the superposition of many generations of aftershocks, however, may be different. In general, the parameter p should be in the range $1.1 \leq p \leq 1.3$.

Finally, we specify the location of each daughter earthquake relative to its parent. There are a wide variety of distributions that we could choose from, but in this paper we assume that a daughter earthquake occurs at a randomly chosen radial distance from the parent earthquake in a randomly chosen direction. Based on results given by FELZER and BRODSKY (2006), we assume a power-law dependence of the radial position in direct analogy to Omori's law. The cumulative distribution function P_{Cr} for the radial distance r_d of each daughter earthquake from the parent earthquake is given by

$$P_{Cr} = \frac{N_d(\geq r_d)}{N_{dT}} = \frac{1}{(1 + r_d/(d \cdot 10^{0.5m_p}))^{q-1}}. \tag{13}$$

The dependence on the magnitude m_p of the parent earthquake introduces a mean radial position of aftershocks that scales with the rupture length of the parent earthquake. The radial position r_d of the daughter earthquake relative to the parent earthquake is selected from this distribution. For each daughter earthquake a random value for P_{Cr} in the range $0 < P_{Cr} < 1$ is generated, and the radial distribution is determined using Eq. (13) in the form

$$r_d = d \cdot 10^{0.5m_p}({P_{Cr}}^{-1/(q-1)} - 1). \tag{14}$$

In order to completely specify the location of the daughter earthquake, its direction relative to the parent earthquake θ_d must be specified. The direction is therefore chosen randomly from the uniform range $0 < \theta_d < 2\pi$.

6. *BASS versus ETAS*

There are many similarities between the BASS model and the ETAS model. More importantly, however, there are also fundamental differences. In considering the ETAS model, we will utilize the formulation given by HELMSTETTER and SORNETTE (2003a). Both

models utilize the concept of multiple orders of aftershocks. The main shock generates a sequence of primary aftershocks; these in turn generate families of secondary aftershocks, and so forth.

The primary difference between the two models is the way in which the number of daughter earthquakes is specified. In the ETAS model, the number of daughter earthquakes produced by a mother earthquake N_{dT} takes the form (SORNETTE and WERNER, 2005b, eq. 3)

$$N_{dT} = k \cdot 10^{\alpha(m_p - m_{\min})}, \tag{15}$$

and the two constants k and α must be specified. In some formulations of ETAS, the mean number n of direct aftershocks per earthquake, averaged over all magnitudes, is used to specify k (SORNETTE and WERNER, 2005b, eq. 4). If Eqs. (4) and (15) are identical, then ETAS is essentially identical to BASS. This is the case if

$$\alpha = b_d \tag{16}$$

and

$$k = 10^{-b_d \Delta m^*}. \tag{17}$$

FELZER *et al*. (2002) argue that Eq. (16) is in fact satisfied. Proponents of ETAS, however, require $\alpha < b_d$ so that n is less than one. The BASS formulation utilizes the modified form of Båth's law as given in Eq. (3) to constrain the number of daughter earthquakes N_{dT} (productivity).

The general ETAS formulation does not satisfy Båth's law. The association of Båth's law to ETAS has been discussed in some detail by HELMSTETTER and SORNETTE (2003c). As shown in their Figure 1, Δm has strong magnitude dependence. In fact, the values of Δm become negative in the vicinity of the minimum magnitude earthquakes $m_{\min}$. In this vicinity, the average largest aftershock is greater than the main shock. This clearly violates scale-invariance and makes results very sensitive to the choice of $m_{\min}$. In the BASS model, the distribution of daughter earthquake magnitudes is fully scale-invariant and is insensitive to the choice of the minimum size earthquake considered.

In terms of the times and positions of daughter earthquakes, BASS and ETAS use identical formulations. Both use Omori's law for times and a modified form of Omori's law for radial positions.

7. *Illustration of the BASS Model*

We first illustrate the principals of the BASS model using a deterministic branching formulation. We begin by considering the distribution of aftershocks associated with a main shock of a prescribed magnitude. The application of Båth's law introduces a characteristic earthquake magnitude Δm, the magnitude difference between the main

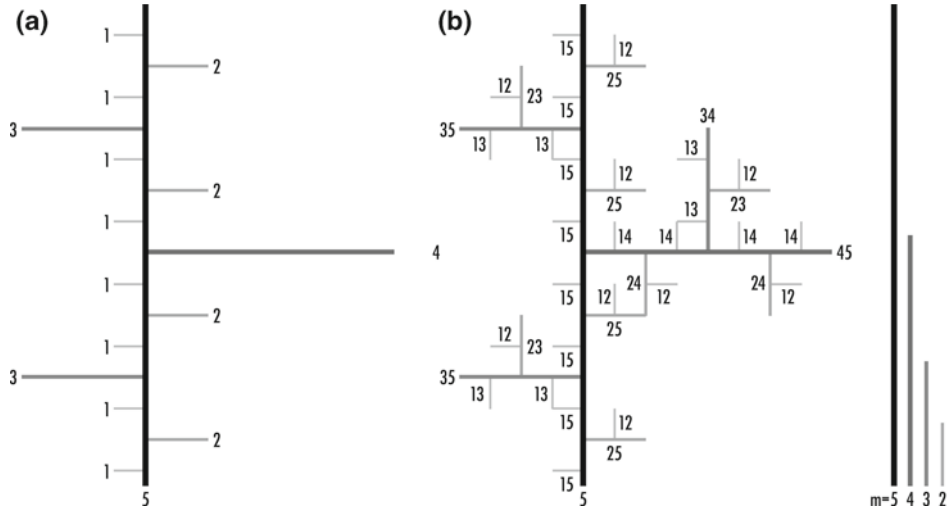


Figure 1
Illustration of our branching model using a discrete set. (a) The primary family of aftershocks is a generator for the fractal construction; (b) the full Tokunaga structure of side-branching aftershocks.

shock and the largest aftershock. For convenience, we take $\Delta m = 1$ and consider the number of earthquakes with unit magnitudes, $m = 1, 2, 3, ...$

Our basic formulation introduces a branching ratio B into the prescription of the number of smaller earthquakes of magnitude i generated by a larger earthquake of magnitude j. The number of daughter earthquakes of magnitude i generated by a parent earthquake of magnitude j, N_{ij}, can be written

$$N_{ij} = B^{j-i-1}. \tag{18}$$

This basic branching is illustrated in Figure 1a for binary branching ($B = 2$). We take the "parent" to have a magnitude $j = 5$. From Eq. (18) we have one daughter earthquake with magnitude $i = 4$, two daughter earthquakes with magnitude $i = 3$, four daughter earthquakes with magnitude $i = 2$, and eight daughter earthquakes with magnitude $i = 1$. That is $N_{45} = 1$, $N_{35} = B = 2$, $N_{25} = B^2 = 4$, and $N_{15} = B^3 = 8$.

TOKUNAGA (1978) introduced the basic branching ratio concept given in Eq. (16) for river network branching. Extensive examples have been given by PECKHAM (1995) and PELLETIER (1999). This branching was also found to be applicable to the structure of diffusion limited aggregation (DLA) clusters (OSSADNIK, 1992), to examples in biology (TURCOTTE *et al.*, 1998), and to clustering (GABRIELOV *et al.*, 1999). Deterministic examples have been given by TURCOTTE and NEWMAN (1996) and by NEWMAN *et al.* (1997). In this paper, we show that this same branching structure is applicable to seismicity.

We extend the basic branching relation given in Eq. (18) to families of aftershocks. That is, we consider the aftershocks of aftershocks. The number of daughter earthquakes

of magnitude i generated by a parent earthquake of magnitude j which was a daughter of a parent earthquake of magnitude k, N_{ijk}, ks given by

$$N_{ijk} = N_{jk}N_{ij}, \tag{19}$$

where N_{ij} is the number of magnitude j aftershocks generated by a parent of magnitude i. Substituting N_{ij} from Eq. (18) into Eq. (19). gives

$$N_{ijk} = N_{jk}B^{j-i-1}. \tag{20}$$

The total number of aftershocks of magnitude i generated by a main shock of magnitude k, N_{ik}, is given by

$$N_{ik} = \sum_{j=i+1}^{k} N_{ijk} = (B+1)^{k-i-1}. \tag{21}$$

The validity of this result can be verified by noting that it gives

$$\begin{aligned} N_{jk} &= 1 \quad \text{if} \quad j = k \\ N_{jk} &= (B+1)^{j-k-1} \quad \text{if} \quad j < k \end{aligned}. \tag{22}$$

Substitution of Eq. (22) into Eq. (21) and carrying out the sum verifies the validity of Eq. (21). The side-branching structure of aftershocks of aftershocks is best illustrated by an example.

The full binary ($B = 2$) side-branching structure of aftershocks for a magnitude 5 main shock is given in Figure 1b. The corresponding numbers N_{ij5} and N_{i5} are given in Table 1. From Eq. (21), we predict $N_{i5} = 3^{4-i}$, which is what we find. A discrete form of the Gutenberg-Richter frequency-magnitude scaling can be written:

$$\log N_{ik} = a - bi. \tag{23}$$

For the sequence given in Table 1, we have $a = \log 81$ and $b = \log 3 = 0.477$, which is low relative to actual aftershock sequences.

Table 1

Illustration of the deterministic BASS model for binary branching $B = 2$, main shock magnitude $m_k = 5$, modified Båth's law $\Delta m = 1$, and minimum magnitude $m_{min} = 1$. The numbers of aftershocks N_{ij5} with magnitude i generated by a parent earthquake of magnitude j are given. The total numbers of aftershocks N_{i5} of magnitude i are also given

Aftershock Magnitude	Parent Earthquake Magnitude				Total N_{ij}
	$j = 5$	$j = 4$	$j = 3$	$j = 2$	
$i = 4$	$N_{455} = 1$				$N_{45} = 1$
$i = 3$	$N_{355} = 2$	$N_{345} = 1$			$N_{35} = 3$
$i = 2$	$N_{255} = 4$	$N_{245} = 2$	$N_{235} = 3$		$N_{25} = 9$
$i = 1$	$N_{155} = 8$	$N_{145} = 4$	$N_{135} = 6$	$N_{125} = 9$	$N_{15} = 27$

Table 2

Illustration of the deterministic BASS model for a branching ratio $B = 9$, main shock magnitude $m_k = 8$, modified Båth's law $\Delta m^* = 1$, and minimum magnitude $m_{min} = 1$. The numbers of aftershocks N_{ij8} with magnitude i generated by a parent earthquake of magnitude j are given. The total number of aftershocks N_{i8} of magnitude i are also given. Note that the total numbers satisfy Gutenberg-Richter frequency-magnitude scaling with $b = 1$

Aftershock Magnitude	Parent Earthquake Magnitude							Total N_{ij}
	$j = 8$	$j = 7$	$j = 6$	$j = 5$	$j = 4$	$j = 3$	$j = 2$	
$i = 7$	1							1
$i = 6$	9	1						10
$i = 5$	81	9	10					100
$i = 4$	729	81	90	100				1000
$i = 3$	6561	729	810	900	1000			10000
$i = 2$	59049	6561	7290	8100	9000	10000		100000
$i = 1$	53144	59049	65610	72900	81000	90000	100000	1000000

As a more realistic example, we take $B = 9$ and a main shock magnitude $k = 8$. The corresponding numbers N_{ij8} and N_{i8} are given in Table 2. We again assume that $\Delta m = 1$ and that $m_{\min} = 1$. The numbers of aftershocks N_{ij8} with magnitude i generated by a parent earthquake of magnitude j are given. The total numbers of aftershocks N_{i8} of magnitude i are also given.

From Eq. (21), we predict $N_{i8} = 10^{7-i}$, which is what we find. From Eq. (23), we find $a = 7$ and $b = 1$ which is a reasonable value for an aftershock sequence. The results given in Table 2 are directly analogous to Table 1. The total families of aftershocks are given. In Table 2 we have $B = 9$ and $m_{ms} = 8$. In Table 1 we have $B = 2$ and $m_{ms} = 5$.

We next determine the entire inventory of earthquakes in a region using our deterministic model. In order to do this, we must specify the magnitude of the largest earthquake in the region, $m_{\max}$. For our example, we take $m_{\max} = 8$. We first specify the numbers of main shocks of magnitude i, N_{im}, using Gutenberg-Richter scaling in the form

$$N_{im} = B^{8-i}. \tag{24}$$

Taking $B = 9$ and $m_{\min} = 1$, the numbers of main shocks, N_{im}, are given in Table 3. Utilizing the results given in Table 2, the numbers of aftershocks N_{ik} with aftershock magnitude i generated by a main shock with magnitude k are also given in Table 3. The total numbers of aftershocks of each magnitude N_{ia} and the total numbers of earthquakes of each magnitude N_{iT} are also given. It is seen that the total number of earthquakes, all main shocks and their aftershocks, satisfy the scaling relation

$$N_{iT} = (B + 1)^{8-i}. \tag{25}$$

The fraction of all earthquakes that are aftershocks increases systematically from 10% at magnitude seven to 52% at magnitude one.

Table 3

Entire inventory of earthquakes in a region given by our deterministic model. We specify the largest earthquake to have $m_{max} = 8$. The numbers of main shocks of magnitude i, N_{im} as obtained from Eq. (24) with $B = 9$ are given. The total numbers of aftershocks of magnitude i generated by a main shock of magnitude k, N_{ik}, are also given as well as the numbers of aftershocks and total numbers of earthquakes of magnitude k

Earthquake Magnitude	Number of Main shocks	Parent Earthquake Magnitude							Total Aftershocks	Total Earthquakes
		$j = 8$	$j = 7$	$j = 6$	$j = 5$	$j = 4$	$j = 3$	$j = 2$		
$i = 8$	1									1
$i = 7$	9	1							1	10
$i = 6$	81	10	9						19	100
$i = 5$	729	100	90	81					271	1000
$i = 4$	6561	1000	900	810	729				3439	10000
$i = 3$	59049	10000	9000	8100	7290	6561			40451	100000
$i = 2$	531441	100000	90000	81000	72900	65610	59049		468559	1000000
$i = 1$	4782969	1000000	900000	810000	729000	656100	590490	531441	5217031	10000000

8. Probabilistic BASS Simulation

We now give a specific probabilistic simulation using the BASS model. A complete stochastic aftershock sequence will be generated. In order to start the simulation, it is necessary to choose a main shock amplitude, m_k. In this example we take $m_k = 7$.

A. We first determine the distribution of primary aftershocks generated by the main shock by taking the main shock to be the parent earthquake.
   1. The first step is to determine the total number of primary aftershocks from Eq. (4). In addition to the parent earthquake magnitude $m_p = m_k = 7$ it is necessary to specify a *b*-value for the daughter earthquakes, we take $b_d = 1$ throughout the simulation, and the modified Båth's law magnitude difference Δm^*, we take $\Delta m^* = 1.0$ for simplicity. It is also necessary to specify the minimum magnitude considered. For this example, we take $m_{\min} = 2$. With these values, we find from Eq. (4) that the total number of primary aftershocks is $N_{dT} = 10^{4.0} = 10000$.
   2. We generate $N_{dT} = 10000$ random numbers for the P_{Cm} in the range $0 < P_{Cm} < 1$, and the magnitudes of the $N_{dT} = 10000$ primary aftershocks are determined using Eq. (7) and the parameter values given above.
   3. We next utilize the generalized form of Omori's law given in Eq. (8) to obtain the time of occurrence of each aftershock. In order to specify the cumulative distribution function P_{Ct} given in Eq. (11), we require the two parameters *c* and *p*. Based on the results given by Yamanaka and Shimazaki (1990), Felzer *et al.* (2003), and Shcherbakov *et al.* (2004), we take $c = 0.1$ days and $p = 1.25$. We again generate $N_{dT} = 10000$ random numbers for P_{Ct} in the range $0 < P_{Ct} < 1$, and the times of occurrence of the $N_{dT} = 10000$ primary aftershocks are determined using Eq. (12) and the parameter values given above. Note that the

times of occurrence of the aftershocks t_d are not correlated with their magnitudes m_d.

4. Finally, we utilize the cumulative distribution function P_{Cr} given in Eq. (13) to specify the radial distance of the daughter earthquakes from the parent earthquake. For primary aftershocks, these distances are from the main shock. To fully specify P_{Cr}, we require two parameters d and q. Based on the results given by FELZER and BRODSKY (2006), we take $d = 4$ m and $q = 1.35$. We generate another $N_{dT} = 10000$ random numbers for P_{Cr} in the range $0 < P_{Cr} < 1$, and the radial distances of the $N_{dT} = 10000$ primary aftershocks are determined using Eq. (14) and the parameter values given above. Again, these radial positions of these aftershocks are not correlated with either their magnitudes or their times of occurrence. A final set of $N_{dT} = 10000$ random numbers are generated in the range $0 < \theta < 2\pi$. The value of θ for each aftershock is taken as the angle of the aftershock relative to some reference direction.

B. Each of the primary aftershocks are next treated as a parent earthquake, and steps A1 to A4 are repeated.
   1. For each primary aftershock, the number of secondary aftershocks is obtained using Eq. (4). This number has a strong dependence on the magnitude of the primary aftershock under consideration. The magnitude of each secondary aftershock is then determined using random numbers and the distribution given in Eq. (7). Note that the magnitudes to be determined do not depend on the magnitude of the parent earthquake (the primary aftershock).
   2. The time of occurrence of each secondary aftershock is then determined using random numbers and the distribution given in Eq. (12). Note that the time of occurrence of each secondary aftershock is the time since the occurrence of the parent earthquake (the primary aftershock).
   3. The radial position of each secondary aftershock is determined using random numbers and the distribution given in Eq. (14). Note that the radial position is relative to the position of the parent earthquake (the primary aftershock). The direction relative to the parent earthquake is also randomly selected. Note also that the parent magnitude in Eq. (14) is the magnitude of the parent primary aftershock.

C. Each secondary aftershock is taken to be a parent earthquake, and a family of daughter second-order aftershocks is generated using the procedure outlined in B1 through B3. The procedure is further repeated to higher orders until no more aftershocks are generated.

The magnitudes of the aftershocks as a function of times of occurrence since the main shock are given in Figure 2. There are 101015 aftershocks in the simulation spanning twenty two generations. Since, as was pointed out, there are 10000 primary aftershocks, this simulation generated 91015 second- and higher-order aftershocks. The magnitude of the largest aftershock in this simulation is $m = 6.6$, thus $\Delta m = 0.4$. Note that the Δm in

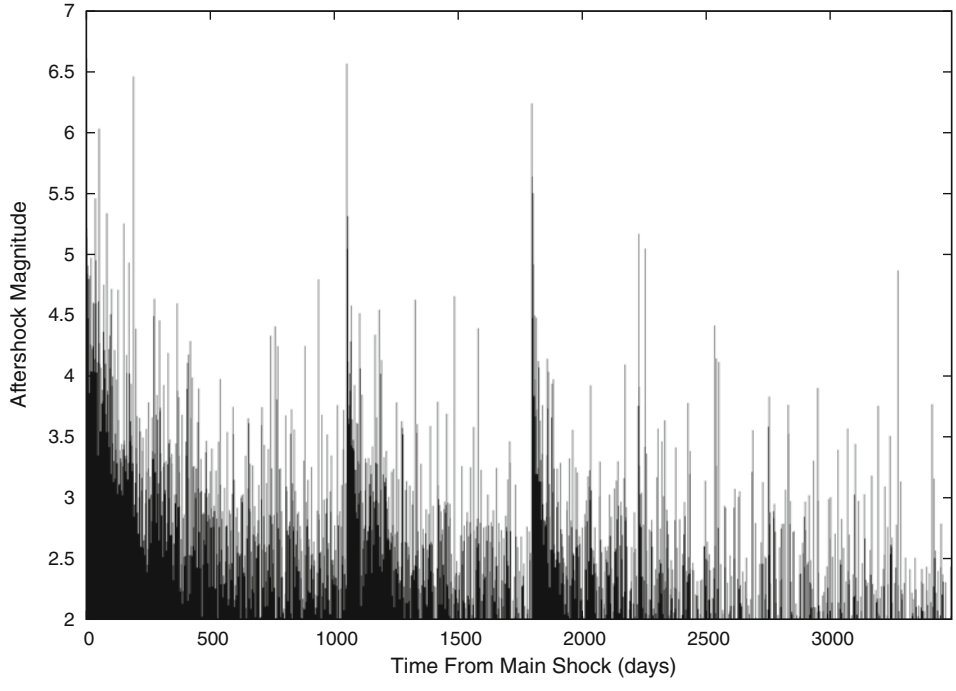


Figure 2
Plot of magnitudes as a function of time (in days) over the first year for the first four generations of an aftershock sequence based on an initial $m = 7.0$ event at time $t = 0$. Note that large aftershocks generate their own aftershocks sequences.

each simulation is different although we take $\Delta m^* = 1$. It is clearly seen in Figure 1 that large primary aftershocks generate their own sequences of higher-order aftershocks.

The positions of the aftershocks relative to the main shock are given in Figure 3. Again, it is clearly seen that clusters of higher-order aftershocks surround the large primary aftershocks. The cumulative Gutenberg-Richter frequency-magnitude statistics of the aftershocks are given in Figure 4. The frequency-magnitude distribution for all aftershocks is well approximated by the Gutenberg-Richter relation (Eq. (1) taking $b = 1$ and $\Delta m^* = 1$.

9. Discussion

Probabilistic seismic hazard assessments play many roles. These include: 1) alerting the public to the level of risk, 2) influencing seismic building codes and seismic retrofitting, 3) setting earthquake insurance premiums, and 4) motivating earthquake hazard preparations. We have discussed two distinct approaches to seismic hazard

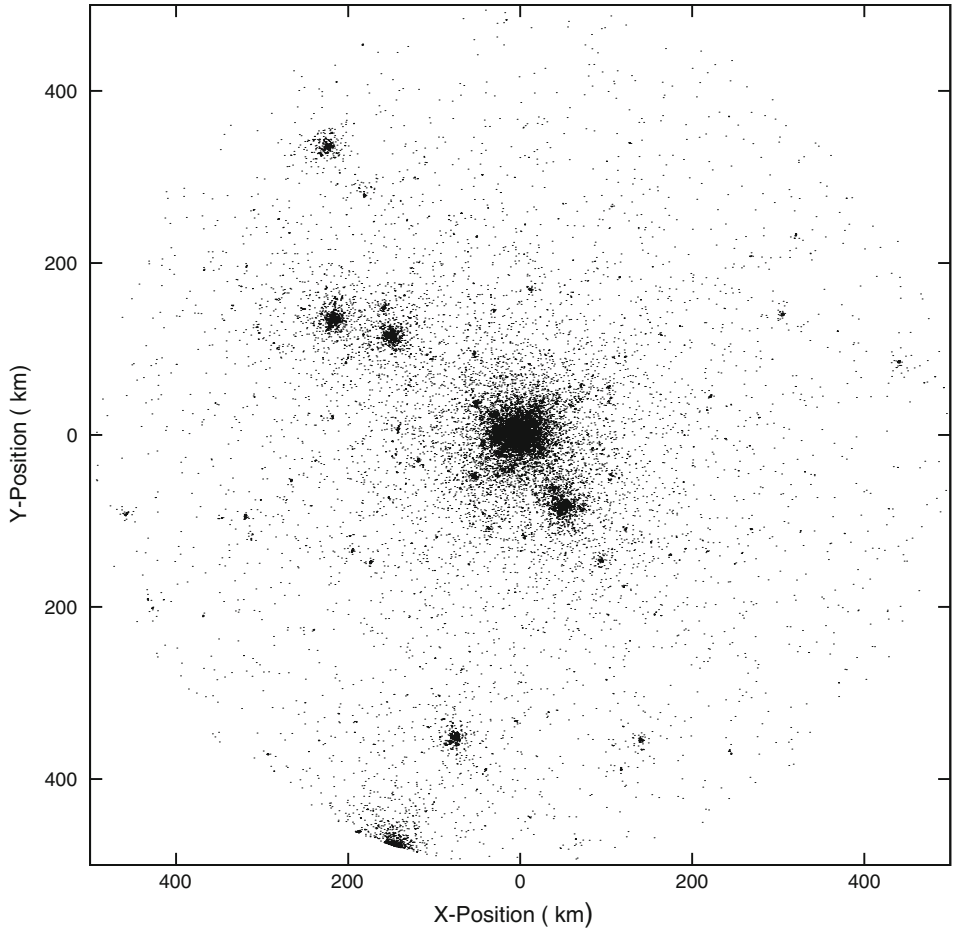


Figure 3
Plot of aftershock positions for the first four generations of an aftershock sequence based on an initial $m = 7.0$ event at location $r = 0$. Note that each generation's aftershocks are clustered about their respective main shocks. Only plotted are aftershocks that fall within a 500 km radius surrounding the main shock.

assessment. The first uses fault-based models. The risk of an earthquake on mapped faults is assessed. This can be done in several ways. The statistics of occurrence of a characteristic earthquake on each fault is prescribed. This requires the magnitude, mean recurrence time, coefficient of variation, and a distribution function for recurrence times. Evidence favors the applicability of the Weibull distribution (YAKOVLEV *et al.*, 2006).This distribution includes Poisson (random) and periodic limits. Fault-based models can be constrained using simulations. Two examples we have discussed are Virtual California and SPEM. These models include a specified tectonic drive and interactions between fault segments.

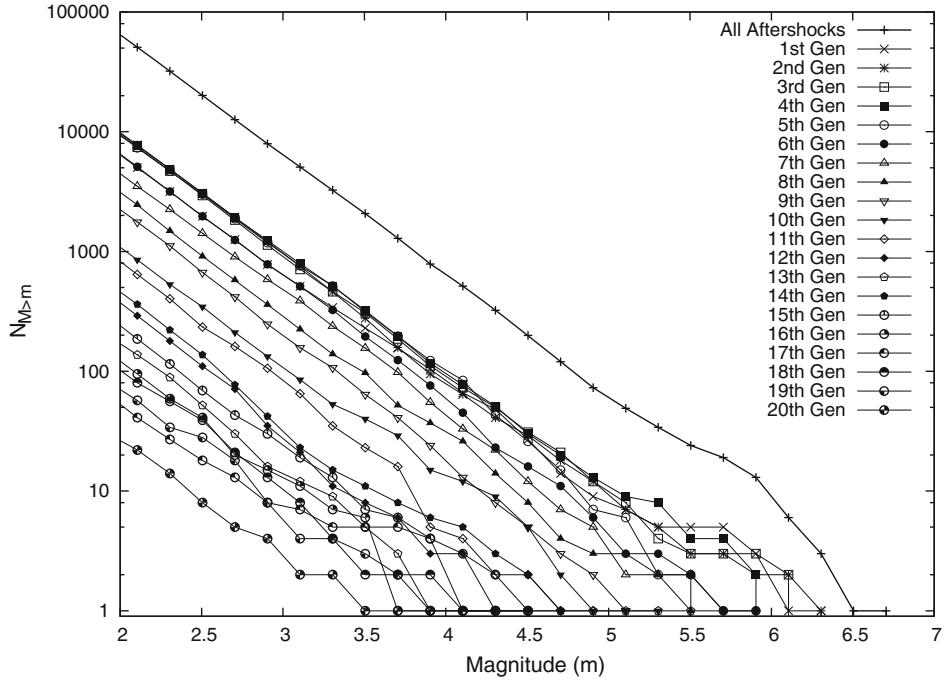


Figure 4
Plot of the Gutenberg-Richter frequency-magnitude relation for an aftershock sequence based on an initial $m = 7.0$ event. Note that this particular realization consisted of twenty two generations (only the first twenty are shown).

A major difficulty with fault-based models is that many large earthquakes occur on faults that have not been mapped. This difficulty can be alleviated by introducing a random background of seismicity. This background, however, should be correlated the regional seismicity.

The second approach to seismic hazard assessment is to use seismicity-based models. Future earthquakes are associated with past earthquakes. The simplest approach is relative intensity (RI) models. The rates of occurrence of small earthquakes (say, $m = 2$) in gridded cells (say, $0.1^\circ \times 0.1^\circ$) are extrapolated to larger magnitudes using Gutenberg-Richter frequency-magnitude scaling.

It is also possible to extrapolate past seismicity forward in time using aftershock models. The ETAS model has been used extensively for this purpose. Each past earthquake can be used as a parent earthquake, and future aftershocks can be determined. In this paper we present the BASS model as an alternative to ETAS. The BASS model is fully self-similar, satisfies all relevant scaling laws, and is simple to implement.

It is clear that there are many alternative approaches to probabilistic seismic hazard assessment. In order to test alternative models for California earthquakes, a competition for Regional Earthquake Likelihood Models (RELM) was sponsored by the Southern

California Earthquake Center (SCEC). Forecasts for $m > 5$ earthquakes during the period 1 January, 2006 to 31 December, 2011 on a $0.1° \times 0.1°$ grid of cells were solicited. Probabilities of occurrence were required for each grid. Nine competing forecasts were submitted and have been summarized by FIELD (2007a). The contrasts between the smeared fault based forecasts and the highly gridded seismicity-based forecasts stands out. Several of the seismicity-based forecasts utilized ETAS models. At the end of the five-year period, the forecasts will each be scored and a winner declared. A direction for future work would be to drive a BASS simulation with a Virtual California simulation. This would combine the fault based and seismicity based approaches.

Acknowledgements

The authors would like to acknowledge valuable discussions of BASS with William Newman and Robert Shcherbakov. This work has been supported through NASA grant NGT5 (JRH), though a US Department of Energy grant to UC Davis DE-FG03-95ER14499 (JRH and JBR), and through NSF grant ATM-0327558 (DLT). The authors would like to thank Ian Main and an anonymous reviewer for substantial comments that significantly improved our paper.

REFERENCES

BÄTH, M. (1965), *Lateral inhomogeneities in the upper mantle*, Tectonophysics *2*, 483–514.

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*, 24359–24372.

BUFE, C.G. and VARNES, D.J. (1993), *Predictive modeling of the seismic cycle of the greater San Francisco Bay region*, J. Geophys. Res. *98*, 9871–9883.

CONSOLE, R. and MURRU, M. (2001), *A simple and testable model for earthquake clustering*, J. Geophys. Res. *106*, 8699–8711.

CONSOLE, R., MURRU, M., and CATALLI, F. (2006), *Physical and stochastic models of earthquake clustering*, Tectonophysics *417*, 141–153.

CONSOLE, R., MURRU, M., and LOMBARDI, A.M. (2003), *Refining earthquake clustering models*, J. Geophys. Res. *108*, 2468.

ELLSWORTH, W.L., MATHEWS, M.V., NADEAU, R.M., NISHENKO, S.P., REASENBERG, P.A., and SIMPSON, R.W. (1999), *A physically-based earthquake recurrence model for estimation of long-term earthquake probabilities*, Open-File Report 99-522, US Geological Survey.

FELZER, K.R., BECKER, T.W., ABERCROMBIE, R.E., EKSTROM, G., and RICE, J.R. (2002) *Triggering of the 1999 m_w 7.1 hector mine earthquake by aftershocks of the 1992 m_w 7.3 landers earthquake*, J. Geophys. Res. *107*(B9), 2190.

FELZER, K.R., ABERCROMBIE, R.E., and EKSTROM, G. (2003), *Secondary aftershocks and their importance for aftershock forecasting*. Bull. Seismol. Soc. Am. *93*(4), 1433–1448.

FELZER, K.R., ABERCROMBIE, R.E., and EKSTROM, G. (2004), *A common origin for aftershocks, foreshocks, and multiplets*, Bull. Seismol. Soc. Am. *94*, 88–98.

FELZER, K.R. and BRODSKY, E.E. (2006), *Decay of aftershock density with distance indicates triggering by dynamic stress*, Nature, *441*, 735–738.

FIELD, E.H. (2007a). *Overview of the working group for the development of regional earthquake likelihood models (RELM)*, Seis. Res. Lett. *78*, 7–16.

FIELD, E.H. (2007b). *A summary of previous working groups on California earthquake probabilities*, Seismol. Soc. Am. Bull. *97*, 1033–1053.

FRANKEL, A.F., MUELLER, C., BARNHARD, T., PERKINS, D., LEYENDECKER, E.V., DICKMAN, N., HANSON, S., and HOPPER, M. (1996), *National seismic hazard maps*, Open-File Report 96-532, US Geological Survey.

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*, 5293.

GERSTERBERGER, M., WIEMER, S., and JONES, L. (2004), *Real-time forecasts of tomorrow's earthquakes in California: A new mapping tool*, Open-File Report 2004-1390, US Geological Survey.

GERSTERBERGER, M.C., WIEMER, S., JONES, L.M., and REASENBERG, P.A. (2005), *Real-time forecasts of tomorrow's earthquakes in California*, Nature *435*, 328–331.

GOES, S.D.B. and WARD, S.N. (1994), *Synthetic seismicity for the San Andreas fault*, Annali Di Geofisica *37*, 1495–1513.

GROSS, S. and RUNDLE, J.B. (1998), *A systematic test of time-to-failure analysis*, Geophys. J. Int. *133*, 57–64.

GUO, Z.Q. and OGATA, Y. (1997), *Statistical relations between the parameters of aftershocks in time, space, and magnitude*, J. Geophys. J. Res. *102*(B2), 2857–2873.

GUTENBERG, B. and RICHTER, C.F., *Seismicity of the Earth and Associated Phenomena* (Princeton University Press, Princeton, NJ 1954).

HELMSTETTER, A. (2003), *Is earthquake triggering driven by small earthquakes?* Phys. Rev. Let. *91*, 0585014.

HELMSTETTER, A. and SORNETTE, D. (2002a), *Diffusion of epicenters of earthquake aftershocks, Omori's law, and generalized continuous-time random walk models*, Phys. Rev. E *66*(6), 061104.

HELMSTETTER, A. and SORNETTE, D. (2002b), *Subcritical and supercritical regimes in epidemic models of earthquake aftershocks*, J. Geophys. Res. *107*(B10), 2237.

HELMSTETTER, A. and SORNETTE, D. (2003a), *Foreshocks explained by cascades of triggered seismicity*, J. Geophys. Rev, *108*(B10), 2457.

HELMSTETTER, A. and SORNETTE, D. (2003b), *Predictability in the epidemic-type aftershock sequence model of interacting triggered seismicity*, J. Geophys. Rev. *108*(B10), 2482.

HELMSTETTER, A.S. and SORNETTE, D. (2003c), *Båth's law derived from the Gutenberg-Richter law and from aftershock properties*, Geophys. Res. Lett. *30*(20), 2069.

HELMSTETTER, A.S. and SORNETTE, D. (2003d), *Importance of direct and indirect triggered seismicity in the ETAS model of seismicity*. Geophys. Res. Lett. *30*(11), 1576.

HELMSTETTER, A.S. and SORNETTE, D., and GRASSO, J.R. (2003a), *Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws*, J. Geophys. Res. *108*(B1, 2046).

HELMSTETTER, A.S., OUILLON, G., and SORNETTE, D. (2003b), *Are aftershocks of large California earthquakes diffusing?* J. Geophys. Res. *108*(B10):2483.

HELMSTETTER, A., HERGARTEN, S., and SORNETTE, D. (2004), *Properties of foreshocks and aftershocks of the non-conservative self-organized critical Olami-Feder-Christensen model*, Phys. Rev. E *70*, 046120.

HELMSTETTER, A., KAGAN, Y.Y., and JACKSON, D.D. (2006). *Comparison of short-term and time-independent earthquake forecast models for southern California*, Bull. Seismol. Soc. Am. *96*, 90–106.

HOLLIDAY, J.R., CHEN, C.C., TIAMPO, K.F., RUNDLE, J.B., TURCOTTE, D.L., and DONNELLAN, A. (2007), *A RELM earthquake forcast based on pattern informatics*, Seis. Res. Lett. *78*(1), 87–93.

HOLLIDAY, J.R., NANJO, K.Z., TIAMPO, K.F., RUNDLE, J.B., and TURCOTTE, D.L. (2005), *Earthquake forecasting and its verification*, Nonlinear Processes in Geophysics, *12*, 965–977.

HOLLIDAY, J.R., RUNDLE, J.B., TIAMPO, K.F., KLEIN, W., and DONNELLAN, A. (2006a), *Modification of the pattern informatics method for forecasting large earthquake events using complex eigenvectors*, Tectonophys. *413*, 87–91.

HOLLIDAY, J.R., RUNDLE, J.B., TIAMPO, K.F., KLEIN, W., and DONNELLAN, A. (2006b), *Systematic procedural and sensitivity analysis of the pattern informatics method for forecasting large*($M \geq 5$) *earthquake events in southern California* Pure Appl. Geophys.

KAGAN, Y.Y. and KNOPOFF, L. (1981), *Stochastic synthesis of earthquake catalogs*, J. Geophys. Res *86*(4), 2853–2862.

KEILIS-BOROK, V.I. (1990), *The lithosphere of the earth as a nonlinear system with implications for earthquake prediction*, Rev. Geophys. *28*, 19–34.

KEILIS-BOROK, V. (2002), *Earthquake predictions: State-of-the-art and emerging possibilities*, An. Rev. Earth Planet. Sci. *30*:1–33.

KEILIS-BOROK, V., SHEBALIN, P., GABRIELOV, A., and TURCOTTE, D. (2004), *Reverse tracing of short-term earthquake precursors*, Phys. Earth Planet. Int. *145*, 75–85.

KOSSOBOKOV, V.G., KEILIS-BOROK, V.I., TURCOTTE, D.L., and MALAMUD, B.D. (2000), *Implications of a statistical physics approach for earthquake hazard assessment and forecasting*, Pure Appl. Geophys. *157*, 2323–2349.

LEPIELLO, E., GODANO, C., and DE ARCANGELIS, L. (2007), *Dynamically scaling in branching models for seismicity* Phys. Rev. Lett. *98*, 098501.

MAIN, I. (1996), *Statistical physics, seismogenesis, and seismic hazard* Rev. Geophys. *34*, 433–462.

MAIN, I.G. (1999), *Applicability of time-to-failure analysis to accelerated strain before earthquakes and volcanic eruptions*, Geophys. J. Int *139*, F1–F6.

NEWMAN, W.I., TURCOTTE, D.L., and GABRIELOV, A.M. (1997), *Fractal trees with side branching*, Fractals *5*, 603–614.

OGATA, Y. (1988), *Statistical models for earthquake occurrences and residual analysis for point processes*, J. Am. Stat. Assoc. *83*, 9–27.

OGATA, Y. (1989), *Statistical model for standard seismicity and detection of anomalies by residual analysis*, Tectonophysics *169*, 159–174.

OGATA, Y. (1992), *Detection of precursory relative quiescence before great earthquakes through a statistical model*, J. Geophys Res. *97*, 19845–19871.

OGATA, Y. (1998), *Space-time point process models for earthquake occurrences*, Ann. Inst. Statist. Math. *50*, 379–402.

OGATA, Y. (1999), *Seismicity analysis through point-process modeling: A review*, Pure Appl. Geophys. *155*, 471–507.

OGATA, Y. (2001a), *Exploratory analysis of earthquake clusters by likelihood-based trigger models*, J. Appl. Probab. *38A*, 202–212.

OGATA, Y. (2001b), *Increased probability of large earthquakes near aftershock regions with relative quiescence*, J. Geophys. Res. *106*, 8729–8744.

OGATA, Y. (2004), *Space-time model for regional seismicity and detection of crustal stress changes*, J. Geophys. Res. *109*, B06308.

OGATA, Y., MATSUURA, R.S., and KATSURA, K. (1993), *Fast likelihood computation of epidemic type aftershock-sequence model*, Geophys. Res. Lett. *20*, 2143–2146.

OGATA, Y., JONES, L.M., and TODA, S. (2003), *When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California*, J. Geophys. Res. *108*, 2318.

OGATA, Y. and ZHUANG, J. (2006), *Space-time ETAS models and an improved extension*, Tectonophysics *413*, 13–23.

OSSADNIK, P. (1992), *Branch order and ramification analysis of large diffusion limited aggregation clusters*, Phys. Rev. A *45*, 1058–1066.

PECKHAM, S.D. (1995), *New results for self-similar trees with applications to river networks*, Water Resour. Res. *31*, 1023–1029.

PELLETIER, J.D. (1999), *Self-organization and scaling relationships of evolving river networks*, J. Geophys. Res. *104*, 7359–7375.

REASENBERG, P.A. (1999), *Foreshock occurrence rates before large earthquake worldwide*, Pure. Appl. Geophys. *155*, 355–379.

REASENBERG, P.A. and JONES, L.M. (1989), *Earthquake hazard after a mainshock in California*, Science *243*(4895), 1173–1176.

RIKITAKE, T., *Earthquake Forecasting and Warning*, (D. Reidel Publishing Co, Dordrecht. 1982)

ROBINSON, R. and BENITES, R. (1995), *Synthetic seismicity models of multiple interacting faults*, J. Geophys. Res. *100*, 18229–18238.

ROBINSON, R. and BENITES, R. (1996), *Synthetic seismicity models for the Wellington Region, New Zealand: Implications for the temporal distribution of large events*, J. Geophys. Res. *101*, 27833–27844.

RUNDLE, J.B., TIAMPO, K.F., KLEIN, W., and MARTINS, J.S.S. (2002), *Self-organization in leaky threshold systems: The influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting*, Proc. Natl. Acad. Sci. U.S.A. *99*, 2514–2521, Suppl. 1.

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.

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 and Space *55*(8), 761–771.

RUNDLE, J.B., RUNDLE, P.B., and DONNELLAN, A. (2005), *A simulation-based approach to forecasting the next great San Francisco earthquake*, Proc. Natl. Acad. Sci. *102*(43), 15363–15367.

RUNDLE, P.B., RUNDLE, J.B., TIAMPO, K.F., DONNELLAN, A., and TURCOTTE, D.L. (2006), *Virtual California: Fault model, frictional parameters, applications*, Pure Appl. Geophys. *163*, 1819–1846.

SAICHEV, A., HELMSTETTER, A., and SORNETTE, D. (2005), *Power-law distributions of offspring and generation numbers in branching models of earthquake triggering*, Pure Appl. Geophys. *162*, 1113–1134.

SAICHEV, A. and SORNETTE, D. (2004), *Anomalous power law distribution of total lifetimes of branching processes: Application to earthquake aftershock sequences*, Phys. Rev. E *70*(4), 046123.

SAICHEV, A. and SORNETTE, D. (2005a), *Distribution of the largest aftershocks in branching models of triggered seismicity: Theory of the universal Båth's law*, Phys. Rev. E *71*(5), 056127.

SAICHEV, A. and SORNETTE, D. (2005b), *Vere-Jones' self-similar branching model*, Phys. Rev. E *72*, 056122.

SAICHEV, A. and SORNETTE, D. (2006a), *Power-law distribution of seismic rates: theory and data analysis*, Eur. Phys. J. *B49*, 377–401.

SAICHEV, A. and SORNETTE, D. (2006b), *Renormalization of branching models of triggered seismicity from total to observed seismicity*, Eur. Phys. J. *B51*, 443–459.

SAICHEV, A. and SORNETTE, D. (2006c), *"Universal" distribution of interearthquake times explained*, Phys. Rev. Lett. *97*, 078501.

SAICHEV, A. and SORNETTE, D. (2007a), *Power-law distributions of seismic rates*, Tectonophysics *431*, 7–13.

SAICHEV, A. and SORNETTE, D. (2007b), *Theory of Earthquake recurrence times*, J. Geophys. Res. *112*, B04313.

SAMMIS, C.G., BOWMAN, D.D., and KING, G. (2004), *Anomalous seismicity and accelerating moment release preceding the 2001-2002 earthquakes in northern Baha California, Mexico*, Pure Appl. Geophys *161*, 2369–2378.

SHCHERBAKOV, R. and TURCOTTE, D.L. (2004), *A modified form of Båth's law*, Bull. Seismol. Soc. Am. *94*, 1968–1975.

SHCHERBAKOV, R., TURCOTTE, D.L., and RUNDLE, J.B. (2004), *A generalized Omori's law for earthquake aftershock decay*, Geophys. Res. Lett. *31*, L11613.

SHCHERBAKOV, R., TURCOTTE, D.L., and RUNDLE, J.B. (2005), *Aftershock statistics*, Pure. Appl. Geophys. *162*, 1051–1076.

SHEBALIN, P., KEILIS-BOROK, V., ZALIAPIN, I., UYEDA, S., NAGAO, T., and TSYBIN, N. (2004), *Advance short-term prediction of the large Tokachi-oki earthquake, September 25, M = 8.1: A case history*, Earth Planets Space *56*, 715–724.

SORNETTE, D. and HELMSTETTER, A. (2002), *Occurrence of finite-time singularities in epedemic models of rupture, earthquakes, and starquakes*, Phys. Rev. Lett. *89*(15), 158501.

SORNETTE, D. and WERNER, M.J. (2005a), *Apparent clustering and apparent background earthquakes biased by undetected seismicity*, J. Geophys. Res. *110*, B09303.

SORNETTE, D. and WERNER, M.J. (2005b), *Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Båth's law, and observed aftershock sequences*, J. Geophys. Res. *110*(B8), B08304.

TIAMPO, K.F., RUNDLE, J.B., MCGINNIS, S., GROSS, S.J., and KLEIN, W. (2002a), *Eigenpatterns in southern California seismicity*, J. Geophys. Res. *107*(B12), 2354.

TIAMPO, K.F., RUNDLE, J.B., MCGINNIS, S., and KLEIN, W. (2002b), *Pattern dynamics and forecast methods in seismically active regions*, Pure Appl. Geophys. *159*, 2429–2467.

TOKUNAGA, E. (1978), *Consideration on the composition of drainage networks and their evolution*, Geographical Rep. Tokya Metro. Univ. *13*, 1–27.

TURCOTTE, D.L., HOLLIDAY, J.R., and RUNDLE, J.B. (2007), *BASS, an alternative to ETAS*, Geophys. Res. Lett. *34*, L12303.

TURCOTTE, D.L. and NEWMAN, W.I. (1996), *Symmetries in geology and geophysics* Proc. Natl. Acad. Sci. *93*, 14295–14300.

TURCOTTE, D.L., PELLETIER, J.D., and NEWMAN, W.I. (1998), *Networks with side branching in biology*, J. Theor. Biol. *193*, 577–592.

UTSU, T. (1984), *Estimation of parameters for recurrence models of earthquakes*, Earthq. Res. Insti.-Univ. Tokyo, *59*, 53–66.

VERE-JONES, D. (1969), *A note on the statistical interpretation of Båth's law*, Bull. Seismol. Soc. Am. *59*, 1535–1541.

VERE-JONES, D. (2005), *A class of self-similar random measure*, Advan. Appli. Probab. *37*, 908–914.

WARD S.N. (1992), *An application of synthetic seismicity in earthquake statistics: The Middle America trench*, J. Geophys. Res. *97*(B5), 6675–6682.

WARD, S.N. (1996), *A synthetic seismicity model for southern California: cycles, probabilities, and hazard*, J. Geophys. Res. *101*(B10), 22393–22418.

WARD, S.N. (2000), *San Francisco Bay Area earthquake simulations: a step toward a standard physical earthquake model*, Bull. Seismol. Soc. Am. *90*(2), 370–386.

WORKING GROUP ON CALIFORNIA EARTHQUAKE PROBABILITIES (1988), *Probabilities of large earthquakes occurring in California on the San Andreas fault*, Open-File Report 88-398, US Geological Survey.

WORKING GROUP ON CALIFORNIA EARTHQUAKE PROBABILITIES (1990), *Probabilities of large earthquakes in the San Francisco Bay region, California*, Circular *1053*, US Geological Survey.

WORKING GROUP ON CALIFORNIA EARTHQUAKE PROBABILITIES (1995), *Seismic hazards in southern California: probable earthquakes, 1994–2024*, *Seis. Soc. Am. Bull*. *85*, 379–439.

WORKING GROUP ON CALIFORNIA EARTHQUAKE PROBABILITIES (2003) *Earthquake probabilities in the San Francisco Bay Region, 2002–2031*, Open-File Report 2003–214, US Geological Survey.

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. Seismol. Soc. Am. *96*, 1995–2007.

YAMANAKA, Y. and SHIMAZAKI, K. (1990), *Scaling relationship between the number of aftershocks and the size of the main shock*, J. Phys. Earth *38*(4), 305–324.

ZHUANG, J. and OGATA, Y. (2006), *Properties of the probability distribution associated with the largest event in an earthquake cluster and their implications to foreshocks*, Phys. Rev. E *73*, 046134.

ZHUANG, J., OGATA, Y., and VERE-JONES, D. (2002), *Stochastic declustering of space-time earthquake occurrences*, J. Am. Stat. Assoc. *97*, 369–380.

ZHUANG, J., OGATA, Y., and VERE-JONES, D. (2004), *Analyzing earthquake clustering features by using stochastic reconstruction*, J. Geophys. Res. *109*, B05301.

(Received April 30, 2007, revised October 26, 2007, accepted October 26, 2007)
Published Online First: July 11, 2008

Pure appl. geophys. 165 (2008) 1025–1047
0033–4553/08/061025–23
DOI 10.1007/s00024-008-0345-5

Pure and Applied Geophysics

A Finite Element Algorithm of a Nonlinear Diffusive Climate Energy Balance Model

R. Bermejo,[1] J. Carpio,[1] J. I. Díaz,[2] and P. Galán del Sastre[3]

Abstract—We present a finite element algorithm of a climate diagnostic model that takes as a climate indicator the atmospheric sea-level temperature. This model belongs to the category of energy balance models introduced independently by the climatologists M.I. Budyko and W.D. Sellers in 1969 to study the influence of certain geophysical mechanisms on the Earth climate. The energy balance model we are dealing with consists of a two-dimensional nonlinear parabolic problem on the 2-sphere with the albedo terms formulated according to Budyko as a bounded maximal monotone graph in $\mathbb{R}^2$. The numerical model combines the first-order Euler implicit time discretization scheme with linear finite elements for space discretization, the latter is carried out for the special case of a spherical Earth and uses quasi-uniform spherical triangles as finite elements. The numerical formulation yields a nonlinear problem that is solved by an iterative procedure. We performed different numerical simulations starting with an initial datum consisting of a monthly average temperature field, calculated from the temperature field obtained from 50 years of simulations, corresponding to the period 1950–2000, carried out by the Atmosphere General Circulation Model HIRLAM.

Key words: Climate, nonlinear energy balance, finite elements.

1. Introduction

During recent decades there has been significant progress in climate modelling with the construction and testing of several Atmosphere-Ocean-General-Circulation-Models. These models are the ultimate tool that can be used to study and predict the Earth's climate system, in that they can include many phenomena taking part in it. However, there remain difficulties for these numerical models to be fully reliable. The first type of difficulty pertains to the lack of understanding of the physical nature of some of these phenomena such as, for example, sub-grid scale processes; so that, they have to be parameterized in order to be included in the models. However, one can argue that most of the sub-grid scale processes can be handled by direct numerical simulation (DNS) of the Navier-Stokes equations, the problem is that in the light of present and near future

[1] Departamento de Matemática Aplicada, Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, C/ José Gutiérrez Abascal 2, 28006 Madrid, Spain. E-mail: rbermejo@etsii.upm.es

[2] Dept. Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain.

[3] Dept. Matemática Aplicada, Universidad Politécnica de Madrid, E. T. S. Arquitectura. Av. Juan de Herrera, 28040 Madrid, Spain.

computer power such an approach is not practical for the moment. A second source of difficulty arises from the computational and numerical resources these models demand to perform well designed experiments; although one may expect that this latter problem can be partially alleviated with the continuous improvements and advances in computer technology, as well as with the development of more accurate and efficient numerical methods. In parallel with the development of general circulation models, the climatologists have developed simpler models intended to clarify the role of some phenomena, whose influence on the evolution of the climate system is considered to be very significant. This approach to the understanding of the climate phenomenology yields the so-called hierarchy of climate models. Perhaps the simplest class of models which may produce interesting results to understand the gross features of the past glacial and interglacial epochs are the so called Energy Balance Models (hereafter, EBMs) which are based on the balance between the incoming solar energy and the energy reflected to the outer space. Although simple in construction, these models may yield under different assumptions to nonlinear problems quite difficult to analyze; this being the reason why these models have caught the attention of many mathematicians. The progress of the mathematical analysis for the EBMs was a function of the different assumptions made on the spatial domain and the nonlinear terms involved in the equation. Among the many results that have appeared in the literature we mention here, in particular, the ones concerning discontinuous co-albedo functions due to XU (1991) and DIAZ (1993) for the one-dimensional case. The analysis of DIAZ (1993) was extended to two dimensions, but with $c(x) \equiv 1$, in DIAZ and TELLO (1999) and HETZER (1990). Many other references can be found in DIAZ (1996).

As for works on the numerical approximation of EBMs, we mention the contributions of North and co-workers such as HYDE *et al*. (1990) and NORTH and COAKLEY (1979), and HETZER *et al*. (1989), where some numerical experiments were carried out. In North and co-workers model the numerical method consists of a first-order Euler implicit scheme for time discretization combined with an spectral method (Legendre polynomial expansion for latitude and trigonometric polynomial expansion for longitude) for space discretization. On the other hand, HETZER *et al*. (1989) use a stationary quasi-linear energy balance model in their study on multiparameter sensitivity analysis of the solutions. In this model, the albedo function is continuous, while the nonlinearity originates from the radiation term which is modelled according to the Stefan-Boltzman radiation law. The model is formulated in spherical coordinates and uses second-order finite differences to discretize the diffusion terms, dealing with the singularities at the poles in an *ad hoc* manner. More recently, BERMEJO *et al*. (2007) formulate and analyze a finite element model of a global nonlinear EBM of Budyko type with a nonlinear diffusion term modelled by the so-called p-Laplacian and a nonlinear discontinuous co-albedo function. The advantages of this finite element model, as well as the model of this paper, are the flexibility to use variable meshes, in particular, if one wants to properly resolve the mushy regions which appear in the transition between ice-covered and ice-free regions, and the form to avoid the singularities at the

poles, which appear when the problem is formulated in spherical coordinates and discretized by grid-point methods such as finite differences, and finite elements of bounded finite volumes.

The layout of the paper is as follows. We introduce in Section 2 the model. In Section 3 we present the mathematical formulation of the model as well as mathematical properties and results concerning the existence and uniqueness of the solution. Section 4 is devoted to the numerical formulation of the model, which is carried out for the special case of a spherical Earth and uses quasi-uniform spherical triangles as finite elements. Finally, Section 5 contains numerical experiments in which we have taken as initial condition a temperature calculated by averaging 50 years of surface temperature data given by the atmospheric general circulation model HIRLAM.

2. The Model

Roughly speaking, the energy balance on the Earth surface is established according to the following law

$$\text{Variation of internal energy} = R_a - R_e + D, \tag{1}$$

where R_a denotes the amount of solar energy absorbed by the earth, R_e is the amount of infrared energy radiated to the space and D is a term which represents the diffusion of heat energy by atmospheric turbulence. Let $u(t, x)$ be the atmospheric sea-level temperature in Celsius degrees, i.e., $u(t, x)$ is defined on $[0, T) \times \mathcal{M}$, where $\mathcal{M}$ is a compact Riemannian manifold without boundary approximating the Earth surface; in fact, $\mathcal{M}$ is a 2-sphere of radius a. Under suitable conditions, the variation of internal energy can be expressed as $c(x)\partial u/\partial t$, where $c(x)$ is the heat capacity (we neglect the possible time dependence of c). The constitutive assumptions for the terms on the right-hand side of (1) are the following:

$$R_a = QS(t, x)\beta(x, u), \tag{2}$$

where Q is the so-called solar constant which is the average (over a year and over the surface of the Earth) value of the incoming solar radiative flux, Q is currently believed to be $Q = \frac{1}{4}(1360\,\mathrm{Wm}^{-2} \pm 2\,\mathrm{Wm}^{-2})$, the function $S(t, x)$ is the normalized seasonal distribution of heat flux entering the top of the atmosphere known as the insolation function. The incident solar flux at the top of the atmosphere at time t and latitude θ can be computed from celestial mechanics (see, e.g., Sellers, 1969); however, we shall use in our model the approximated formulas derived from the exact Sellers formulas by North and Coakley (1979). Specifically, in our model

$$S(t, x) = S_0(t) + S_1(t)\sin\theta + S_2(t)\left(\frac{3\sin^2\theta - 1}{2}\right), \tag{3a}$$

with

$$\begin{cases} S_0(t) = 1 + 2e\cos(2\pi t - \lambda), \\ S_1(t) = S_1[\cos 2\pi t + 2e\sin\lambda\sin 2\pi t], \\ S_2(t) = S_2[1 + 2e\cos(2\pi t - \lambda)], \end{cases} \tag{3b}$$

where θ is the latitude of the point $x \in \mathcal{M}$, e denotes the eccentricity of the earth 's orbit, presently, $e = 0.017$; λ is the angle formed by the lines connecting the Sun with the position of the Earth at the Northern Hemisphere winter solstice and the perihelion, at present $\lambda = -20°$; so that, the perihelion occurs shortly after the winter solstice in the Northern Hemisphere. The coefficients S_1 and S_2 depend upon the obliquity, δ, the present value of δ is 23.45° , so that $S_1 = -0.796$ and $S_2 = -0.477$. The unit of time t is 1 year, with $t = 0$ corresponding to the Northern Hemisphere winter solstice.

The term $\beta(x, u)$ is the so-called co-albedo function that takes values between 0 and 1. $\beta(x, u)$ represents the ratio between the absorbed solar energy and the incident solar energy at the point x on the Earth surface; obviously, $\beta(x, u)$ depends on the nature of the Earth surface. For instance, it is well known that on ice sheets $\beta(x, u)$ is considerably smaller than on the ocean surface because the white color of the ice sheets reflects a large portion of the incident solar energy, whereas the ocean, due to its dark color and high heat capacity, is able to absorb a larger amount of the incident solar energy. We further distinguish between ocean ice sheets and land ice sheets in our model. Following the approach of BUDYKO (1969) we take $\beta(x, u)$ as a nonlinear discontinuous function of the spatial coordinates x and the temperature u of the form given by GRAVES *et al.* (1993):

$$\beta(x, u) = a_0 + a_1\sin\theta + a_2\left(\frac{3\sin^2\theta - 1}{2}\right) + a_I(u), \tag{4}$$

where the coefficients a_0, a_1 and a_2 may depend on time and represent the background albedo characterizing the U-shaped dependence of the albedo. The coefficient a_I takes care of the changes of the albedo in the presence of snow cover and is a function of the temperature u. Table 1, borrowed from GRAVES *et al.* (1993), shows the average values of a_0, a_1 and a_2 calculated from the monthly values of these parameters tabulated in Table 1 of GRAVES *et al.* (1993)

The values of a_I (u) are displayed in Table 2.

Notice that $\beta(x, u)$ is only discontinuous at the level sets $u = u_{s1}$ and $u = u_{s2}$, with $u_{s1} = -2$°C or -5°C and $u_{s2} = -7$°C or -12°C, due to the fact that a_I (u) is

Table 1

Coefficients of the co-albedo function

	Average Sky	Clear Sky
a_0	0.679	0.848
a_1	− 0.012	− 0.020
a_2	− 0.241	− 0045

Table 2

The values of $a_I(u)$

	Average Sky	Clear Sky
$a_{I(Land)}$ (u)	-0.14 if $u < -2°$C, $-[0.14, 0.0]$ if $u = -2°$C, 0.0 otherwise,	-0.50 if $u < -5°$C, $-[0.50, 0.0]$ if $u = -5°$C, 0.0 otherwise,
$a_{I(Ocean)}$(u)	-0.07 if $u < -7°$C, $[-0.07, 0.0]$if $u = -7°$C, 0.0 otherwise,	-0.25 if $u < -12°$C, $[-0.25, 0.0]$if $u = -12°$C, 0.0 otherwise,

discontinuous at these sets. Moreover, $\beta(x, u)$ is nonlinear because $a_I(u)$ is. To see that this statement is true, we must recall the definition of a linear function; that is, if $a_I(u)$ were a linear function then it would follow that given u_1 and u_2 and the real parameters l_1 and l_2, $a_I(l_1u_1 + l_2u_2) = l_1a_I(u_1) + l_2a_I(u_2)$, but it is obvious from the definition of $a_I(u)$ that the latter equality does not hold. Hence, $a_I(u)$ is a nonlinear function. It is worth remarking that $\beta(x, u)$ is not a single-valued function, rather, since for $u = u_{s1}$ (resp. $u = u_{s2}$) $a_I(u) \in [-0.14,0.0]$ or $a_I(u) \in [-0.5,0.0]$ (resp. $a_I(u) \in [-0.07,0.0]$ or $a_I(u) \in [-0.25,0.0]$) then for these values of u the only thing we know is that $\beta(x, u)$ is in bounded real intervals, but we do not know which points of these intervals are $\beta(x, u)$; this is the reason why we say that $\beta(x, u)$ is a multi-valued relation, or by abuse of mathematical language, it is said that $\beta(x, u)$ is a multi-valued graph. So that, it makes sense to write $z \in \beta(x, u)$ as we do below.

The term $R_e(u)$ was modelled by Budyko by performing a linear regression fitting to empirical data as

$$R_e(u) = Bu + C, \tag{5}$$

where B and C are empirical parameters relating the outgoing infrared flux to the surface temperature. According to GRAVES *et al*. (1993) the values that fit best the observations in a least square sense are shown in Table 3.

As for the diffusion term D, Budyko and Sellers proposed the expression

$$D = div(k(x)\nabla u),$$

where $k(x)$ is an eddy diffusion coefficient given by the formula (GRAVES *et al*., 1993):

Table 3

Coefficients of Budyko radiation energy Re(u) = Bu + C

	Average Sky	Clear Sky
$C(\mathrm{Wm}^{-2})$	212.8	249.8
$B(\mathrm{Wm}^{-2\circ}\ C^{-1})$	1.9	2.26

$$k(x) = k_0(1 + k_1 \sin^2\theta + k_2 \sin^4\theta). \tag{6}$$

The coefficients k_0, k_1 and k_2 are given in Table 4.

Finally, we show the values per unit area of the heat capacity $c(x)$. This coefficient is assumed to be a piecewise continuous function, depending on whether the local surface is land, ice or sea. See Table 5.

By substituting the above expressions into (1) we obtain the following energy balance model:

$$(P)\begin{cases} c(x)u_t - div_{\mathcal{M}}(k(x)\nabla_{\mathcal{M}}u) + Bu + C \in QS(t,x)\beta(x,u) & \text{in} \quad (0,T)\times\mathcal{M} \\ u(0,x) = u_0(x) & \text{on} \quad \mathcal{M}, \end{cases}$$

where the initial datum u_0 always will be assumed to be bounded. More precise structural assumptions to solve (P) are formulated in Section 3. A special feature of (P) is that the presence of the co-albedo function $\beta(x,\ u)$ may be responsible, in the case of a discontinuous function, of both the existence of free boundaries at the level sets u_{s1} and u_{s2}, and multiple solutions for certain initial conditions (even if the problem is formulated in terms of a parabolic type equation).

3. On the Existence and Uniqueness of Solutions of the Model (P)

To state the mathematical formulation of (P) we need to recall some basic concepts of differential geometry because the spatial domain $\mathcal{M}$ is the 2-sphere of radius a. Given an index set Λ and $\lambda \in \Lambda$, let W_λ be an open subset of $\mathcal{M}$ such that $\{W_\lambda\}_{\lambda\in\Lambda}$ is an open covering of $\mathcal{M}$, and $w_\lambda : W_\lambda \to w_\lambda(W_\lambda) \subset \mathbb{R}^2$ a homeomorphism. For $\lambda \in \Lambda$, the pair $\{W_\lambda, w_\lambda\}$ is called a chart of $\mathcal{M}$ and the family of charts $\{W_\lambda, w_\lambda\}_{\lambda\in\Lambda}$ is called an atlas of $\mathcal{M}$. Given a point $P \in W_\lambda \subset \mathcal{M}$, we set $w_\lambda(P) = (w_\lambda^1(P), w_\lambda^2(P)) = (\theta_\lambda, \varphi_\lambda) \in \mathbb{R}^2$. The

Table 4

Coefficients of the eddy diffusion coefficient

	Average Sky	Clear Sky
k_0	1.1175	1.331
k_1	− 0.957	− 2.258
k_2	0	1.616

Table 5

Heat capacity coefficient values

c_{water} ($\mathrm{Wm}^{-2} \circ C^{-1} year$)	9.7
c_{land} ($\mathrm{Wm}^{-2\circ} C^{-1} year$)	0.016
c_{ice} ($\mathrm{Wm}^{-2\circ} C^{-1} year$)	0.10

tangent space at P is denoted by $T_P\mathcal{M}$. $T_P\mathcal{M}$ is a vector space of dimension 2 with a basis formed by the vectors $\mathbf{e_1} := \partial/\partial\theta_\lambda, \mathbf{e_2} := \partial/\partial\varphi_\lambda$. The tangent bundle $T\mathcal{M}$ is defined as $T\mathcal{M} := \cup_{P\in\mathcal{M}} T_P\mathcal{M}$. A Riemannian metric g on $\mathcal{M}$ is defined from a family of scalar products $g_P : T_P\mathcal{M} \times T_P\mathcal{M} \to \mathbb{R}$.

For a differentiable function $u : \mathcal{M} \to \mathbb{R}$ the tangent gradient $\nabla_\mathcal{M} u \in T\mathcal{M}$, and for $v : \mathcal{M}\to\mathcal{T}$ differentiable, the surface divergence $div_\mathcal{M} v \in \mathbb{R}$. We denote by $L^2(\mathcal{M})$ the set $\{u : \mathcal{M} \to \mathbb{R}$ measurable : $\int_\mathcal{M} |u|^2 dA < \infty\}$. This set is a Hilbert space with inner product

$$(u, v) = \int_\mathcal{M} uvdA$$

and norm

$$\|u\|_{L^2(\mathcal{M})} = \left(\int_\mathcal{M} |u|^2 dA\right)^{1/2}.$$

Analogously,

$$L^2(T\mathcal{M}) = \{X : \mathcal{M} \to T\mathcal{M} \text{ measurable} : \int_\mathcal{M} < X, X > dA < \infty\}.$$

Also, we shall use the spaces $L^\infty(\mathcal{M})$ and $L^\infty(T\mathcal{M})$ defined as

$$L^\infty(\mathcal{M}) = \{u : \mathcal{M} \to \mathbb{R} \text{ measurable} : \operatorname{ess\,sup}_\mathcal{M} |u(x)| < \infty\}$$

and

$$L^\infty(T\mathcal{M}) = \{X : \mathcal{M} \to T\mathcal{M} \text{ measurable} : \operatorname{ess\,sup}_\mathcal{M} |X(x)| < \infty\},$$

where ess sup is a shorthand notation for the essential supremum defined as

$$\operatorname{ess\,sup}_\mathcal{M} |u(x)| = \inf\left\{\sup_{x\in\mathcal{S}} |u(x)| : \mathcal{S} \subset \mathcal{M}, \text{ with } \mathcal{M}\backslash\mathcal{S} \text{ of measure zero}\right\}.$$

We also need the Sobolev space

$$H^1(\mathcal{M}) = \{u \in L^2(\mathcal{M}) : \nabla_\mathcal{M} u \in L^2(T\mathcal{M})\},$$

with inner product

$$((u, v)) = \int_\mathcal{M} uvdA + \int_\mathcal{M} < \nabla_\mathcal{M} u, \nabla_\mathcal{M} v > dA$$

and norm

$$\|u\|_{H^1(\mathcal{M})} = \sqrt{((u, v))}.$$

$H^1(\mathcal{M})$ is the closure of the set of infinitely continuous functions, $C^\infty(\mathcal{M})$, in the H^1-norm. When m integer, $m > 1$, the Sobolev space of order m is the closure of $C^\infty(\mathcal{M})$ in the norm

$$\|u\|_{H^m(\mathcal{M})} = \left(\int_{\mathcal{M}} \left(\sum_{1 \le k \le m} \sum_{i_j=1,2} \sum_{j=1,..,k} |D_{i_1} D_{i_2} ... D_{i_k} u|^2 + |u|^2 \right) dA \right)^{1/2},$$

where $D_1 = D_{\mathbf{e}_1}$ and $D_2 = D_{\mathbf{e}_2}$. When $m = 0, H^0(\mathcal{M}) = L^2(\mathcal{M})$.

Given a bounded and strictly positive function $c(x)$, and $Q > 0$, we consider the problem (P)

$$(P) \begin{cases} c(x)u_t - \operatorname{div}_{\mathcal{M}}(k(x)\nabla_{\mathcal{M}} u) + Bu + C \in QS(t,x)\beta(x,u) + f(t,x) & \text{in } (0,T) \times \mathcal{M}, \\ u(0,x) = u_0(x) & \text{on } \mathcal{M}, \end{cases}$$

under the following assumptions:

(A1) $\beta(x,\cdot)$ is a bounded maximal monotone graph of $\mathbb{R}^2$,
(A2) $f \in L^\infty((0,T) \times \mathcal{M})$,
(A3) $S : [0,T] \times \mathcal{M} \to \mathbb{R}, S \in C^1([0,T] \times \mathcal{M}), 0 < S_0 \le S(t,x) \le S_1$ *a.e.* $x \in \mathcal{M}$, for any $t \in [\,0,T]$,
(A4) $c \in L^\infty(\mathcal{M}), c(x) \ge c_0 > 0$,
(A5) $k \in C(\mathcal{M}), k(x) \ge k_0 > 0$,
(A6) $u_0 \in L^\infty(\mathcal{M})$,
(A7) $B > 0$ and $C > 0$ constants.

Note the presence of a forcing term $f(t, x)$ in the general statement of problem (P). We do not expect the existence of classical solutions to (P) due to the possible discontinuity of the co-albedo function. For this reason, we need the notion of weak solution to (P).

Definition 1. *A function* $u \in C([0,T]; L^2(\mathcal{M})) \cap L^\infty((0,T) \times \mathcal{M})) \cap L^2(0,T; H^1)$ *is termed a bounded weak solution of* (*P*) *if there exists* $z \in L^\infty((0,T) \times \mathcal{M}), z(t,x) \in \beta(x, u(t,x))$ *a.e.* $(t,x) \in (0,T) \times \mathcal{M}$ *such that*

$$\begin{aligned} \int_{\mathcal{M}} c(x)u(T,x)v(T,x)dA - \int_0^T \int_{\mathcal{M}} c(x)v_t(t,x)u(t,x)dAdt \\ + \int_0^T \int_{\mathcal{M}} <k(x)\nabla_{\mathcal{M}} u, \nabla_{\mathcal{M}} v > dAdt + \int_0^T \int_{\mathcal{M}} (Bu + C)vdAdt \\ = \int_0^T \int_{\mathcal{M}} QS(t,x)z(t,x)vdAdt + \int_0^T \int_{\mathcal{M}} fvdAdt \\ + \int_{\mathcal{M}} c(x)u_0(x)v(0,x)dA, \end{aligned} \tag{7}$$

$\forall v \in L^2(0,T;H^1)$ *such that* $v_t \in L^2(0,T;H^{-1})$. *Here* H^{-1} *denotes the dual space of* H^1.

The main results on the existence and uniqueness of bounded weak solutions to problem (P) are collected in Theorem 2 and Theorem 4; the proofs of which can be found in Bermejo *et al*. (2007).

Theorem 2. *Under the above assumptions there exists at least one bounded weak solution of* (*P*). *Moreover, if* $u_0 \in H^1$ *then* $u_t \in L^2(0,T;L^2(\mathcal{M}))$ *and* $\mathrm{div}(k(x)\nabla_{\mathcal{M}}u) \in L^2(0,T;\, L^2(\mathcal{M}))$.

Since $\beta(x, u)$ is considered to be a multi-valued graph discontinuous at the level sets $u = u_{s_1}$ and $u = u_{s_2}$, then there are cases for which problem (P), although parabolic, does not have a unique solution. Nevertheless, it can be proved, Bermejo *et al*. (2007), the uniqueness of the bounded weak solution to (P) in the class of non-degenerate functions which is introduced next.

Definition 3. *Let* $u \in L^\infty(\mathcal{M})$. *Given* ε_0, $0 < \varepsilon_0 < 1$, *for* $\varepsilon \in (0, \varepsilon_0)$ *and* $i = 1,2$ *let*

$$B_{s_i}(u, u_{s_i}; \epsilon) = \{x \in \mathcal{M} :\ |u - u_{s_i}| < \epsilon\}$$

and

$$B_{w_i}(u, u_{s_i}; \epsilon) = \{x \in \mathcal{M} :\ 0 < |u - u_{s_i}| < \epsilon\}.$$

It is said that u is a non-degenerate function in a strong (*resp. weak*) *sense if it satisfies the following strong* (*resp. weak*) *non-degeneracy property: There exists a constant* $C > 0$ *such that for any* $\varepsilon \in (0, \varepsilon_0)$

$$area(B_{s_i}(u, u_{s_i}; \epsilon)) \le C\epsilon \qquad (resp.area(B_{w_i}(u, u_{s_i}; \epsilon)) \le C\epsilon).$$

Theorem 4. *Let* $u_0 \in L^\infty(\mathcal{M})$. *Then:*

(i) *If a bounded weak solution u*(*t*) *to (P) is a strong non-degenerate function for all* $t \in [0,T]$, *then u is the unique bounded weak solution to (P).*

(ii) *For any* $t \in (0, T]$ *there is at most one bounded weak solution u*(*t*) *to (P) in the class of weak non-degenerate functions.*

4. The Numerical Model

4.1. Preliminaries

We now proceed to formulate the numerical method to compute the bounded weak solution to problem (P). This method consists of a combination of $C^0 -$ finite elements for space discretization with a first-order Euler implicit scheme to discretize in time. This time scheme is also used in Hyde *et al*. (1990). We must point out that we choose the Euler implicit scheme for the main reason that our codes have been developed to integrate problem (P) when the diffusion term is also a nonlinear term modelled by the so called p-Laplacian, that is, as $\mathrm{div}_{\mathcal{M}}(|\nabla u|^{p-2}\nabla u)$, p integer > 2; and according to theoretical results of Barret and Liu (1994), and Ju (2000), one may conclude that the optimal time discretization scheme (optimality must be understood here in the sense that there is a balance between computational cost versus accuracy) combined with linear

finite elements to integrate the time dependent p-Laplacian diffusion equation is the first-order Euler implicit scheme. However, we are aware that for problem (P), in which the diffusion terms are linear, it would be more convenient, as one of the reviewers has pointed out to us, to use in combination with finite elements the second- order implicit BDF2 (see Chap. III in HAIRER *et al*., 1993) because the good properties this scheme has for stiff problems. The 2-sphere $\mathcal{M}$ is partitioned into quasi-uniform spherical triangles using the scheme of BAUMGARDNER and FREDERICKSON (1985), which consists of taking as the initial partition D_0 the spherical icosahedron and then to generate a sequence of partitions D_k, $k = 1, 2,...$, by joining the mid-points on the sides of the triangles of the partition $D_{k\text{-}1}$. This procedure yields triangles with the following properties. Let N_k be the number of triangles in the partition D_k, then (a) $\mathcal{M} = \cup_{j=1}^{N_k} T_j, T_j \subset \mathcal{M}$; (b) for $i \neq j, T_i \cap T_i$ is either empty or has one vertex x_p, or T_i and T_j share a common edge γ_{ij}; (c) there exists a positive constant μ such that for all T_j, $h_j/\rho_j < \mu$, where h_j denotes the diameter of T_j and ρ_j is the diameter of the largest circle inscribed in T_j.

Following the approach of DZIUK (1988) to solve by finite elements the Poisson equation on manifolds, it is convenient to view the spherical triangles of the partition D_k of $\mathcal{M}$ as the radial projection onto $\mathcal{M}$ of 2-simplices $\Omega_j \subset \mathbb{R}^3$, such that if T_j is the image of Ω_j, then for all $j, T_j \cap \Omega_j = \{x_{1j}, x_{2j}, x_{3j}\}$, where x_{ij}, $i = 1, 2, 3$, are the vertices of both T_j and Ω_j. By analogy with the elements T_j, the simplices Ω_j form a partition D_{hk} of a polyhedron $\mathcal{M}_h$ such that

$$\mathcal{M}_h := \cup_j \Omega_j,\ \Omega_j \in D_{hk}.$$

We show in Figure 1 the initial icosahedron and the partition D_h after four refinements.

The radial projection is defined as

$$\phi : \mathcal{M}_h \to \mathcal{M}$$

$$\begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \end{pmatrix} \to \begin{pmatrix} \dfrac{a\widehat{x}_1}{\sqrt{(\widehat{x}_1)^2+(\widehat{x}_2)^2+(\widehat{x}_3)^2}} \\ \dfrac{a\widehat{x}_2}{\sqrt{(\widehat{x}_1)^2+(\widehat{x}_2)^2+(\widehat{x}_3)^2}} \\ \dfrac{a\widehat{x}_3}{\sqrt{(\widehat{x}_1)^2+(\widehat{x}_2)^2+(\widehat{x}_3)^2}}. \end{pmatrix},$$

so that, we write

$$\mathcal{M} = \cup_j \phi(\Omega_j)$$

and denote the restriction of ϕ on the element Ω_j by ϕ_j. Note that ϕ is a C^m-diffeomorphism, $m \geq 1$. We define the family of finite element spaces associated with the partitions D_{hk}.

$$\widehat{V}_h = \{\widehat{v}_h \in C^0(\mathcal{M}_h) : \widehat{v}_h|_{\Omega_j} \in P_1(\Omega_j), 1 \leq j \leq N_k\},$$

where $P_1(\Omega_j)$ is the set of polynomials of degree ≤ 1 defined on Ω_j. Let M be the global number of vertices in the partition D_k, and let $\{\alpha_l\}_{l=1}^{M}$ be the set of global basis functions for $\widehat{V}_h$, such that $\alpha_l \in \widehat{V}_h$ and at the vertex $\widehat{x}_j$ $\alpha_l(\widehat{x}_j) = \delta_{jl}$; any $\widehat{v} \in \widehat{V}_h$ can be expressed as

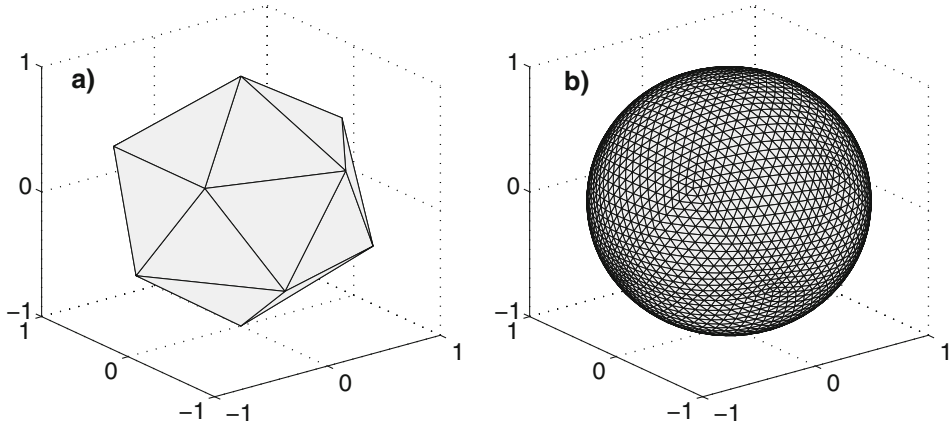


Figure 1
Initial Icosahedron and mesh after 4 refinements.

$$\widehat{v}_h(\widehat{x}) = \sum_{l=1}^{M} \widehat{v}_h(x_l)\alpha_l(\widehat{x}).$$

We define a finite element space $V_h \subset H^1(\mathcal{M})$ associated with the partition D_k via the radial $\phi -$ lifting as follows:

$$V_h = \{v_h \in C^0(\mathcal{M}) : v_h|_{T_j} = \widehat{v}_h \circ \phi_j^{-1} \quad \text{with} \quad \widehat{v}_h \in \widehat{V}_h\}.$$

The approximation spaces V_h and $\widehat{V}_h$ satisfy:
For all $u \in L^\infty((0,T) \times \mathcal{M}) \bigcap L^2(0,T;V)$, $u_t \in L^2((0,T) \times \mathcal{M})$

$$\lim_{h \to 0} \inf_{u_h \in V_h} \|u - u_h\|_{L^\infty((0,T)\times\mathcal{M})} = 0.$$

Moreover, from computational and numerical analysis points of view it is convenient to define the spaces $H^l(\mathcal{M}_h), l \geq 0$ (with the convention that for $l = 0, H^l(\mathcal{M}_h) \equiv L^2(\mathcal{M}_h)$) as

$$H^l(\mathcal{M}_h) = \{\widehat{v} : \mathcal{M}_h \to \mathbb{R} : \textit{for a.e. } x \in \mathcal{M} \textit{ and } v \in H^l(\mathcal{M}), \widehat{v} \circ \phi^{-1}(x) = v(x)\}.$$

In relation with the radial projection ϕ defined on $\mathcal{M}_h$ we have the following results (Bermejo *et al.*, 2007):

Proposition 5. *Let J_{ϕ_j} and $J_{\phi_j^{-1}}$ denote the absolute values of the Jacobian determinants of the mappings ϕ_j and ϕ_j^{-1}, respectively. Then, for h sufficiently small there exist constants C_1 and C_2 independent of h such that*

$$\max_j \|J_{\phi_j} - 1\|_{L^\infty(\Omega_j)} \leq C_1 h^2 \textit{ and } \max_j \|J_{\phi_{j^{-1}}} - 1\|_{L^\infty(T_j)} \leq C_2 h^2.$$

Proposition 6. *For* $1 \leq p \leq \infty$ *there exist constants* c_1 *and* c_2 *such that*

$$c_1\|\widehat{v}\|_{L^p(\mathcal{M}_h)} \leq \|v\|_{L^p(\mathcal{M})} \leq c_2\|\widehat{v}\|_{L^p(\mathcal{M}_h)},$$
$$c_1\|\widehat{v}\|_{H^1(\mathcal{M}_h)} \leq \|v\|_{H^1(\mathcal{M})} \leq c_2\|\widehat{v}\|_{H^1(\mathcal{M}_h)},$$
$$|\widehat{v}|_{H^2(\mathcal{M}_h)} \leq c_2\Big(|v|_{H^2(\mathcal{M})}+h|v|_{H^1(\mathcal{M})}\Big).$$

The relevance of these results, in particular Proposition 6, lies in the fact that by virtue of it the approximation error in the family of finite element spaces $\widehat{V}_h$ is of the same order as the error in the family of spaces V_h associated to the partition D_k of spherical triangles. In terms of the numerical calculations this means that one can substitute the spherical triangles (curved triangles) by plane triangles in $\mathbb{R}^3$ and, therefore, make use of the finite element technology for plane triangles. At this point, we must say that the idea of approximating the 2-sphere by a $\mathbb{R}^3$ polyhedra of triangular faces has a long tradition in numerical computations of atmospheric flows. Just to cite a few, we mention the works of Sadourny *et al*. (1968), and Williamson (1968) at the end of the sixties of the past century, and more recently the integration of the shallow water equations via a Lagrange-Galerkin method carried out by Heinze and Hense (2002), and Giraldo and Warburton (2005).

Since the numerical solution to problem (P) is computed at a discrete set of time instants t_n, with $n = 0, 1, ..., N$, we choose a fixed time step Δt, such that for all n, $t_{n+1} = t_n + \Delta t$, and consider the discrete set $I_N = \{0, t_1, t_2, ..., t_N = T\}$. The numerical solution to (P) is thus the map $U{:}I_N \to V_h$ such that there exists $Z^n \in L^\infty(\mathcal{M}) \bigcap V_h, Z^n \in \beta(x, U^n)$, verifying that for any $v_h \in V_h$

$$(P_{h,\Delta t})\begin{cases} \int_{\mathcal{M}} c\,\frac{U^n-U^{n-1}}{\Delta t} v_h dA + \int_{\mathcal{M}} \langle k\nabla_{\mathcal{M}} U^n, \nabla_{\mathcal{M}} v_h\rangle dA + \\ \int_{\mathcal{M}} (BU^n + C)v_h dA = \int_{\mathcal{M}} QS^n Z^n v_h dA + \int_{\mathcal{M}} f^n v_h dA, \end{cases}$$

where the notation $b(t_n,x) = b^n$ is used unless otherwise stated.

An important property of the finite element space V_h is that if a function $w_h \in V_h$ is an approximation to a function $w \in L^\infty((0,T)\times\mathcal{M})$ that belongs to the class of non-degenerate functions (either strong or weak), then for h sufficiently small w_h also belongs to that class. Specifically, we have the following results. For $i = 1$ and 2, let $B_{s_i}(w, u_{s_i}; \epsilon)$ and $B_{w_i}(w, u_{s_i}; \epsilon)$ be the sets introduced in Section 3, and we consider the level sets

$$A_i = \{x \in \mathcal{M} : w(t,x) = u_{s_i}\}, \qquad A_{hi} = \{x \in \mathcal{M} : w_h(t,x) = u_{s_i}\},$$
$$M_i^{\pm} = \{x \in \mathcal{M} : w(t,x) \gtrless u_{s_i}\} \quad \text{and} \quad M_{hi}^{\pm} = \{x \in \mathcal{M} : w_h(t,x) \gtrless u_{s_i}\}.$$

Note that $\mathcal{M} = A_i \cup M_i^+ \cup M_i^- = A_{hi} \cup M_{hi}^+ \cup M_{hi}^-$. It is easy to ascertain that for $z \in \beta(x,w)$ and $z_h \in \beta(x, w_h)$ it holds

$$\begin{cases} |z - z_h| \leq max|a_I(u)| & \text{if} \quad x \in A_i \cup A_{hi} \cup (M_i^+ \cap M_{hi}^-) \cup (M_i^- \cap M_{hi}^+), \\ |z - z_h| = 0 & \text{if} \quad x \in (M_i^+ \cap M_{hi}^+) \cup (M_i^- \cap M_{hi}^-). \end{cases}$$

Moreover, the following lemma can be proved (Bermejo *et al*. 2007):

Lemma 7. *Given a function* $v \in L^\infty((0,T)\times M)\cap L^2(0,T;V)$, *and its approximation* $v_h \in V_h$, *for h depending on* $\in$ *sufficiently small the relation*

$$A_i \cup A_{hi} \cup (M_i^+ \cap M_{hi}^-) \cup (M_i^- \cap M_{hi}^+) \subset B_{s_i}(v, u_{s_i}; \epsilon)$$

holds for $i = 1$ *and* 2. *Consequently, there exists a constant* $C > 0$ *such that*

$$area\big(A_i \cup A_{hi} \cup (M_i^+ \cap M_{hi}^-) \cup (M_i^- \cap M_{hi}^+)\big) \le C\epsilon.$$

We are now in a condition to state the result on the existence and uniqueness of the solution $\{U^n\}_1^N$ to problem $(P_{h,\Delta t})$, whose proof is given in Bermejo *et al*. (2007).

Lemma 8. *For all* $n = 1,..., N$, *there exists a solution* $U^n \in V_h$ *to problem* $(P_{h,\Delta t})$ *which is unique in the class of strong* (*resp. weak*) *non-degenerate functions*.

An important issue when calculating a numerical solution to a model is to estimate the rate of convergence of the approximate solution to the exact one. Again, appealing to the numerical analysis employed in Bermejo *et al*. (2007) to prove its Theorem 3, we can establish the rate of convergence of U^n to $u(t_n,x)$ for all n.

Theorem 9. *Let* $u(t,x)$ *be the unique non-degenerate bounded weak solution to problem (P), with* $u \in L^2(0,T;H^2(\mathcal{M}))$. *Let* $\{U^n\}_{n=1}^N$ *be the unique solution to problem* $(P_{h,\Delta t})$ *such that for* $n = 1, 2,..., N$ *and* $t \in (t_{n-1}, t_n]$ *we define*

$$U(t) = \frac{t - t_{n-1}}{\Delta t} U^n - \frac{t_n - t}{\Delta t} U^{n-1}.$$

Then, for Δt *and h depending on* $\in$ *being sufficiently small, there exists a constant* $C > 0$ *independent of* Δt *and h such that*

$$\|u - U\|^2_{L^\infty(0,T;L^2(\mathcal{M}))} \le C(\epsilon + \Delta t^2 + h^2) \tag{8}$$

4.2. The Finite element Solution

To calculate the numerical solution we recast problem $(P_{h,\Delta t})$ as follows:
Given the initial condition $U^0 \in V_h$, *for* $n = 1,..., N$, *find* $U^n \in V_h$ *such that for* $v_h \in V_h$

$$\begin{cases} \int_{\mathcal{M}} cU^n v_h dA + \Delta t \int_{\mathcal{M}} \langle k\nabla_{\mathcal{M}} U^n, \nabla_{\mathcal{M}} v_h\rangle dA + \Delta t \int_{\mathcal{M}} (BU^n + C)v_h dA = \\ \int_{\mathcal{M}} cU^{n-1} v_h dA + \Delta t \int_{\mathcal{M}} QS^n Z^n v_h dA + \Delta t \int_{\mathcal{M}} f^n v_h dA, \end{cases} \tag{9}$$

where $Z^n \in L^\infty(\mathcal{M}) \bigcap V_h, Z^n \in \beta(x, U^n)$. Since U^n is unknown so is Z^n, which has to be calculated in the process of determining the solution U^n. To do so we use the following iterative procedure:

Let $Tol \in \mathbb{R}_+, 0 < Tol \ll 1$; *for all* $n = 1, \ldots, N$, *set* $W^0 = U^{n-1}$ *and do:*
for $k = 1, 2,...$
pick up $Z^{n,k-1} \in \beta(x, W^{k-1}), Z^{n,k-1} \in V_h$ *and solve*

$$\begin{cases} \int_{\mathcal{M}} cW^k v_h dA + \Delta t \int_{\mathcal{M}} \langle k \nabla_{\mathcal{M}} W^k, \nabla_{\mathcal{M}} v_h \rangle dA + \Delta t \int_{\mathcal{M}} (BW^k + C) v_h dA = \\ \int_{\mathcal{M}} cU^{n-1} v_h dA + \Delta t \int_{\mathcal{M}} QS^n Z^{n,k-1} v_h dA + \Delta t \int_{\mathcal{M}} f^n v_h dA, \ \forall v_h \in V_h. \end{cases} \tag{10}$$

Stop when

$$\frac{\left\| W^k - W^{k-1} \right\|_{L^2(\mathcal{M})}}{\left\| W^0 \right\|_{L^2(\mathcal{M})}} \leq Tol$$

and then set

$$U^n(x) = W^k(x).$$

By applying the same ideas of the proof of Theorem 1 of Carl (1992) one can prove that this iterative procedure converges to U^n when $k \to \infty$.

To find out the numerical solution W^k, and therefore U^n, we approximate the triangulated 2-sphere $\mathcal{M}$ by the polyhedron $\mathcal{M}_h$ and setting, $\widehat{c}(\widehat{x}) = c \circ \phi(\widehat{x})$, $\widehat{k}(\widehat{x}) = k \circ \phi(\widehat{x})$ and $\widehat{f^n}(\widehat{x}) = f^n \circ \phi(\widehat{x})$, solve instead of (10) the following problem defined on $\mathcal{M}_h$:

For $n = 1, 2, .., N$ *do:*
$\widehat{W}^0(\widehat{x}) = \widehat{U}^{n-1}(\widehat{x})$
for $k = 1,2,...$
pick up $\widehat{Z}^{n,k-1} \in \beta(\widehat{x}, \widehat{W}^{k-1}), \widehat{Z}^{n,k-1} \in \widehat{V}_h$ *and find* $\widehat{W}^k \in \widehat{V}_h$, *such that for* $\widehat{v}_h \in \widehat{V}_h$

$$\begin{cases} \int_{\mathcal{M}_h} \widehat{c} \widehat{W}^k \widehat{v}_h dA_h + \Delta t \int_{\mathcal{M}_h} \widehat{k} \nabla_{\mathcal{M}_h} \widehat{W}^k \cdot \nabla_{\mathcal{M}_h} \widehat{v}_h dA_h + \Delta t \int_{\mathcal{M}_h} (B\widehat{W}^k + C) \widehat{v}_h dA_h = \\ \int_{\mathcal{M}_h} \widehat{c} \widehat{U}^{n-1} \widehat{v}_h dA_h + \Delta t \int_{\mathcal{M}_h} Q\widehat{S}^n \widehat{Z}^{n,k-1} \widehat{v}_h dA_h + \Delta t \int_{\mathcal{M}_h} \widehat{f}^n \widehat{v}_h dA_h. \end{cases} \tag{11a}$$

Stop when

$$\frac{\left\| \widehat{W}^k - \widehat{W}^{k-1} \right\|_{L^2(\mathcal{M}_h)}}{\left\| \widehat{W}^0 \right\|_{L^2(\mathcal{M}_h)}} \leq Tol$$

and set

$$\widehat{U}^n(\widehat{x}) = \widehat{W}^k(\widehat{x}), \tag{11b}$$

and for $x \in \mathcal{M}$ *and* $\widehat{x} \in \mathcal{M}_h$, *such that* $x = \phi(\widehat{x})$,

$$U^n(x) = \widehat{U}^n(\widehat{x}). \tag{11c}$$

Next, we shall describe the method to implement $\nabla_{\mathcal{M}_h} \widehat{u}_h(\widehat{x})$ for any $\widehat{u}_h(\widehat{x}) \in \widehat{V}_h$. Following Dziuk (1988) we write the tangent gradient $\nabla_{\mathcal{M}} u \in L^2(T\mathcal{M})$ when $u \in H^1(\mathcal{M})$ as

$$\nabla_{\mathcal{M}} u = \nabla u - (\overrightarrow{n}_{\mathcal{M}} \cdot \nabla u)\overrightarrow{n}_{\mathcal{M}},$$

where $\overrightarrow{n}_{\mathcal{M}}$ is the unit outward normal vector on $\mathcal{M}$ and $\nabla u = (\partial u/\partial x_i)_{i=1,2,3}$ denotes the gradient of u considered as a function of the Cartesian coordinates (x_1, x_2, x_3) referred to the Cartesian coordinate system, the origin of which is at the center of the sphere. Recalling that for $\widehat{x} \in \mathcal{M}_h$, $\widehat{u}(\widehat{x})$ is a lifting of $u(x)$, i.e., $x \in \mathcal{M}$ is such that $x = \phi(\widehat{x})$ and $\widehat{u}(\widehat{x}) = u \circ \phi(\widehat{x})$, then $\nabla_{\mathcal{M}} u$ will be numerically approximated by the approximation to $\nabla_{\mathcal{M}_h} \widehat{u}(\widehat{x}) \in L^2(T\mathcal{M}_h)$, the expression of which is

$$\nabla_{\mathcal{M}_h} \widehat{u}(\widehat{x}) = \nabla \widehat{u}(\widehat{x}) - (\overrightarrow{n}_{\mathcal{M}_h} \cdot \nabla \widehat{u}(x))\overrightarrow{n}_{\mathcal{M}_h} \text{ for any } \widehat{x} \in \mathcal{M}_h,$$

where $\overrightarrow{n}_{\mathcal{M}_h}$ denotes the unit outward normal vector on $\mathcal{M}_h$, which is a constant vector on each triangular face Ω_j of $\mathcal{M}_h$, defining thus a piecewise constant approximation to $\overrightarrow{n}_{\mathcal{M}}$. $\widehat{u}(\widehat{x})$ is approximated by $\widehat{u}_h(\widehat{x}) \in \widehat{V}_h$ satisfying $\widehat{u}_h(P) \mid_{\Omega_j} \in P_1(\Omega_j)$; that is

$$\widehat{u}_h(\widehat{x}) \mid_{\Omega_j} = \sum_{m=1}^{3} \widehat{U}_m \lambda_m(\widehat{x}),$$

where $\widehat{U}_m = \widehat{u}_h(\widehat{x}_m)$, and the local basis functions $\{\lambda_m(\widehat{x})\}_{m=1}^3$ are the so-called barycentric coordinates defined by the relations

$$\begin{array}{ll} \sum_{m=1}^3 \widehat{x}_{mi}\lambda_m = \widehat{x}_i, & \text{for } i = 1,2,3, \\ \sum_{m=1}^3 \lambda_m = 1 & \forall P \in \Omega_j, \end{array}$$

here $\widehat{x}_i$ are the coordinates of any point $\widehat{x} \in \Omega_j$ and $\widehat{x}_{mi}$ are the coordinates of the vertices of Ω_j. Then, denoting by $\overrightarrow{n}_j$ the unit normal vector on Ω_j we have that for any $\widehat{x} \in \Omega_j$

$$\nabla_{\mathcal{M}_h} \widehat{u}_h(\widehat{x}) = \sum_{m=1}^{3} \widehat{U}_m \nabla \lambda_m - \left(\sum_{l=1}^{3} n_{jl} \sum_{m=1}^{3} \widehat{U}_m \frac{\partial \lambda_m}{\partial \widehat{x}_l} \right) \overrightarrow{n}_j.$$

We notice that by construction of the family of finite element spaces V_h, $\widehat{U}_m$ are also the values $u_h(x_m)$, with $x_m = \phi(\widehat{x}_m)$ being the vertices of the spherical triangles. Moreover, via the local basis functions $\{\lambda_m(\widehat{x})\}$ of the elements Ω_j we can define a set of global basis functions $\{\alpha_l(\widehat{x})\}_{l=1}^M$ for the finite element space $\widehat{V}_h$ that is characterized by the following properties: (1) For each $l, \alpha_l(\widehat{x}) \in \widehat{V}_h$; (2) for $1 \le i, l \le M, \alpha_l(\widehat{x}_i) = \delta_{il}$; (3) for $1 \le j \le N_k$, $1 \le l \le M$ and $1 \le m \le 3$, the restriction of $\alpha_l(\widehat{x})$ on the element Ω_j, i.e., $\alpha_l(\widehat{x}) \mid_{\Omega_j} = \lambda_m(\widehat{x})$ if the mesh node $\widehat{x}_l$ coincides with the m-th vertex of the Ω_j. By properties (1) and (2) the global basis functions $\alpha_l(\widehat{x})$ are piecewise linear polynomials of compact support and each element $\widehat{u}_h(\widehat{x}) \in \widehat{V}_h$ is expressed as

$$\widehat{u}_h(\widehat{x}) = \sum_{l=1}^{M} \widehat{U}_l \alpha_l(\widehat{x}).$$

By property (3) we can evaluate the domain integrals in (11a) as the sum of element integrals using the local basis functions $\{\lambda_m\}$.

Now, we calculate the integral $\int_{\mathcal{M}_h} \widehat{k}\nabla_{\mathcal{M}_h}\widehat{u}_h^n \cdot \nabla_{\mathcal{M}_h}\widehat{v}_h dA_h$ as

$$\int_{\mathcal{M}_h} \widehat{k}\nabla_{\mathcal{M}_h}\widehat{u}_h^n \cdot \nabla_{\mathcal{M}_h}\widehat{v}_h dA_h = \sum_{j=1}^{N_k} \int_{\Omega_j} \widehat{k}\nabla_{\mathcal{M}_h}\widehat{u}_h \cdot \nabla_{\mathcal{M}_h}\widehat{v}_h dA_h, \tag{12a}$$

where the element integral

$$\int_{\Omega_j} \widehat{k}\nabla_{\mathcal{M}_h}\widehat{u}_h \cdot \nabla_{\mathcal{M}_h}\widehat{v}_h dA_h = \widehat{V} S_j \widehat{U}^T,$$

with $\widehat{V} = (\widehat{V}_1, \widehat{V}_2, \widehat{V}_3)$, $\widehat{U} = (\widehat{U}_1, \widehat{U}_2, \widehat{U}_3)$, $\widehat{V}_k$ and $\widehat{U}_k$ being the values of $\widehat{V}_h$ and $\widehat{u}_h$ at the vertices of Ω_j respectively, and S_j is the Ω_j-element symmetric matrix the entries of which are

$$s_{ik} = \int_{\Omega_j} \widehat{k}\nabla\Psi_i \cdot \nabla\lambda_k dA_h = \int_{\Omega_j} \widehat{k}(\nabla\lambda_i - (\overrightarrow{n}_j \cdot \nabla\lambda_i)\overrightarrow{n}_j) \cdot \nabla\lambda_k dA_h, \quad 1 \le i,k \le 3. \tag{12b}$$

Note that s_{ik} are the entries of the stiffness matrix corresponding to the two-dimensional Laplace operator minus $\int_{\Omega_j}(\overrightarrow{n}_j \cdot \nabla\lambda_i)(\overrightarrow{n}_j \cdot \nabla\lambda_k)dA_h$. We are now in a condition to describe how the evaluation of integrals of (11a) yields an algebraic system of equations the solution of which is formed by the values of $\widehat{W}^k$ at the vertices of the spherical triangles.

$$\int_{\mathcal{M}_h} \widehat{k}\nabla_{\mathcal{M}_h}\widehat{W}^k \cdot \nabla_{\mathcal{M}_h}\widehat{v}_h dA_h = \sum_{j=1}^{N_k} \int_{\Omega_j} \widehat{k}\nabla_{\mathcal{M}_h}\widehat{W}^k \cdot \nabla_{\mathcal{M}_h}\widehat{v}_h dA_h = \widehat{\mathbf{V}}^T \mathbf{S}\widehat{\mathbf{W}}^k,$$

$$\int_{\mathcal{M}_h} (\widehat{c}\widehat{W}^k + \Delta t B\widehat{W}^k)\widehat{v}_h dA_h = \sum_{j=1}^{N_k} \int_{\Omega_j} (\widehat{c}\widehat{W}^k + \Delta t B\widehat{W}^k)\widehat{v}_h = \widehat{\mathbf{V}}^T(\mathbf{M}_1 + \Delta t B\mathbf{M}_2)\widehat{\mathbf{W}}^k,$$

$$\Delta t C \int_{\mathcal{M}_h} \widehat{v}_h dA_h = \Delta t C \sum_{j=1}^{N_k} \int_{\Omega_j} \widehat{v}_h dA_h = \Delta t C \widehat{\mathbf{V}}^T \mathbf{L},$$

$$\int_{\mathcal{M}_h} \widehat{c}\widehat{U}^{n-1}\widehat{v}_h dA_h = \sum_{j=1}^{N_k} \int_{\Omega_j} \widehat{c}\widehat{U}^{n-1}\widehat{v}_h dA_h = \widehat{\mathbf{V}}^T \mathbf{M}_1 \widehat{\mathbf{U}}^{n-1},$$

$$\Delta t Q \int_{\mathcal{M}_h} \widehat{S}^n \widehat{Z}^{n,k-1}\widehat{v}_h dA_h = \Delta t Q \sum_{j=1}^{N_k} \int_{\Omega_j} \widehat{S}^n \widehat{Z}^{n,k-1}\widehat{v}_h dA_h = \Delta t Q \widehat{\mathbf{V}}^T \cdot \widehat{\mathbf{Z}}^{n,k-1},$$

$$\Delta t \int_{\mathcal{M}_h} \widehat{f}^n \widehat{v}_h dA_h = \Delta t \sum_{j=1}^{N_k} \int_{\Omega_j} \widehat{f}^n \widehat{v}_h dA_h = \Delta t \widehat{\mathbf{V}}^T \cdot \mathbf{F}^n.$$

In these formulas the M-dimensional vector $\widehat{\mathbf{V}}^T := (\widehat{V}_1, \ldots \widehat{V}_M)$, $\widehat{V}_i$ being the value of $\widehat{V} \in \widehat{V}_h$ at the mesh point x_i. Similarly, $\widehat{\mathbf{W}}^k := (\widehat{W}_1^k, \ldots \widehat{W}_M^k)^T$, $\widehat{\mathbf{W}}^{k-1} := (\widehat{W}_1^{k-1}, \ldots \widehat{W}_M^{k-1})^T$ and $\widehat{\mathbf{U}}^{n-1} := (\widehat{U}_1^{n-1}, \ldots \widehat{U}_M^{n-1})^T$. **S,** $\mathbf{M}_1$ and $\mathbf{M}_2$ are sparse symmetric $M \times M$ matrices obtained by assembling the corresponding element matrices. Thus,

$$\mathbf{S} = \bigcup_{j=1}^{N_k} S_j$$

where S_j is the element matrix whose entries are given by (12b).

$$\mathbf{M}_1 = \bigcup_{j=1}^{N_k} M_{1j}, \mathbf{M}_2 = \bigcup_{j=1}^{N_k} M_{2j} \quad \text{and} \quad \mathbf{L} = \bigcup_{j=1}^{N_k} L_j$$

where M_{1j}, M_{2j} and L_j are element matrices with entries

$$\begin{cases} m_{1ik} = \int_{\Omega_j} \widehat{c}\lambda_i\lambda_k dA_h, \\ m_{2ik} = \int_{\Omega_j} \lambda_i\lambda_k dA_h, \qquad (1 \le i,k \le 3). \\ l_{ii} = \int_{\Omega_j} \lambda_i dA_h, \qquad l_{ij} = 0 \text{ when } i \ne j \end{cases}$$

The M-dimensional vector $\widehat{\mathbf{Z}}^{n,k-1} = (\widehat{Z}_1^{n,k-1}, \ldots, \widehat{Z}_M^{n,k-1})^T$ is obtained by assembling the element vectors $\widehat{\mathbf{Z}}_j^{n,k-1}$:

$$\widehat{\mathbf{Z}}^{n,k-1} = \bigcup_{j=1}^{N_k} \widehat{\mathbf{Z}}_j^{n,k-1},$$

the entries of $\widehat{\mathbf{Z}}_j^{n,k-1}$ being given by

$$\widehat{z}_l^{n,k-1} = \int_{\Omega_j} \widehat{S}^n \widehat{Z}^{n,k-1} \lambda_l dA_h, \quad 1 \le l \le 3.$$

Likewise, the vector $\mathbf{F}^n = (F_1^n, \ldots, F_M^n)^T$ is obtained by assembling of the element vectors $\mathbf{F}_j^n$ the entries of which are the values of the integrals

$$\int_{\Omega_j} \widehat{f}^n \lambda_k dA_h, \quad 1 \le k \le 3.$$

We use the 7 points Hammer quadrature rule for triangles, which is exact for polynomials of degree 5, to calculate the integrals because the expressions for $S(t,x), \widehat{Z}^{n,k-1}$ and $\widehat{k}(\widehat{x})$ give integrands that are polynomials of degree 4.

Important features that make this formulation attractive for computations are the absence of the so-called "pole problem" and the discretization of the Laplace-Beltrami operator (i.e., the Laplace operator defined on an $(d-1)$-dimensional manifold in $\mathbb{R}^d$) can be managed with the computer codes developed for the Laplace operator in a Cartesian coordinate system.

The algebraic version of the iteration algorithm is then:

Iteration algorithm (*algebraic version*)

For $n = 1, 2, \ldots, N$ *do:*

$\widehat{\mathbf{W}}^0 = \widehat{\mathbf{U}}^{n-1}$

for $k = 1, 2, \ldots$ *do:*

calculate

$$\widehat{Z}^{n,k-1} \in \beta(\widehat{x}, \widehat{W}^{k-1}), \widehat{Z}^{n,k-1} \in \widehat{V}_h$$

and

$$\widehat{\mathbf{Z}}^{n,k-1} = \int_{\mathcal{M}_h} \widehat{S}^n \widehat{Z}^{n,k-1} \widehat{v}_h dA_h \quad \forall \widehat{v}_h \in \widehat{V}_h,$$

then solve

$$\widehat{\mathbf{W}}^k = \mathbf{M}_1 \widehat{\mathbf{U}}^{n-1} + \Delta t(Q\widehat{\mathbf{Z}}^{n,k-1} + \mathbf{F}^n) - \Delta t C\mathbf{L}. \tag{13}$$

Stop when

$$\frac{\left\| \widehat{\mathbf{W}}^k - \widehat{\mathbf{W}}^{k-1} \right\|_{l^2}}{\left\| \widehat{\mathbf{W}}^0 \right\|_{l^2}} \leq Tol$$

and set

$$\begin{cases} \widehat{W}^k(\widehat{x}) = \sum_{l=1}^{M} \widehat{W}_l^k \alpha_l(\widehat{x}) \\ \widehat{\mathbf{U}}^n = \widehat{\mathbf{W}}^k \end{cases} \tag{14}$$

5. Numerical Experiments

Starting with an initial condition that we may consider representative of the present climate temperature, we shall run our model to predict the seasonal evolution of the surface temperature as well as the influence of the concentration of CO_2 on the increase of such a temperature. All the numerical experiments are performed under the hypothesis of *average sky* and with the co-albedo coefficients a_0, a_1 and a_2 being piecewise monthly constants; the values of which are borrowed from Table 1 of GRAVES *et al.* (1993) .

The initial condition is obtained by averaging for every month of the year the surface temperature data given by the general circulation model HIRLAM from the year 1950 up to the year 2000. Figure 2 represents the distribution of the initial temperature which corresponds to December. The computational mesh consists of 20480 triangles and 10242 mesh points, which means an average $h = 0.0431$ rads $\simeq 260$ Kms. We calculate the numerical solution taking a time step length $\Delta t = 0.01 = 3.6$ days, and solving (13) with a tolerance of 0.001. Since $S(t, x)$ depends periodically on time with a period of one year, then after an initial transient state the solution of the model will also be periodic because the coefficients of our model do not depend on time (e.g., BADII and DIAZ 1999). This can be seen in Figure 7 where we represent the evolution of the temperature at a point near Madrid (Spain) under different concentrations of CO_2 in the atmosphere, see equation (15) below.

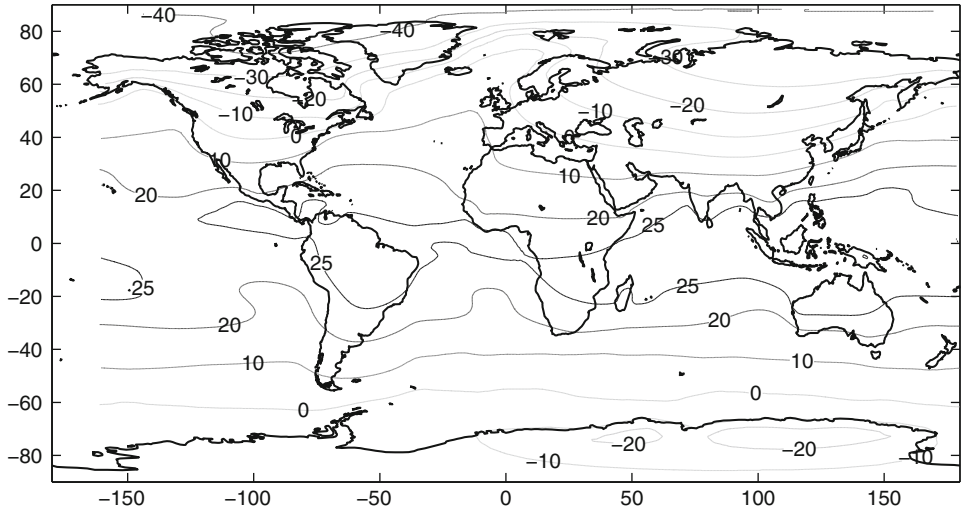


Figure 2
Distribution of temperature at time $t = 0$.

We have noted that the transient period of the model, also known as the spin-up period, is about 9 years. After this, the solution becomes periodic with a period of about 1 year as long time numerical experiments (40 years) have shown. It seems that this periodic state is stable for the parameters used in our calculations. This is the reason we

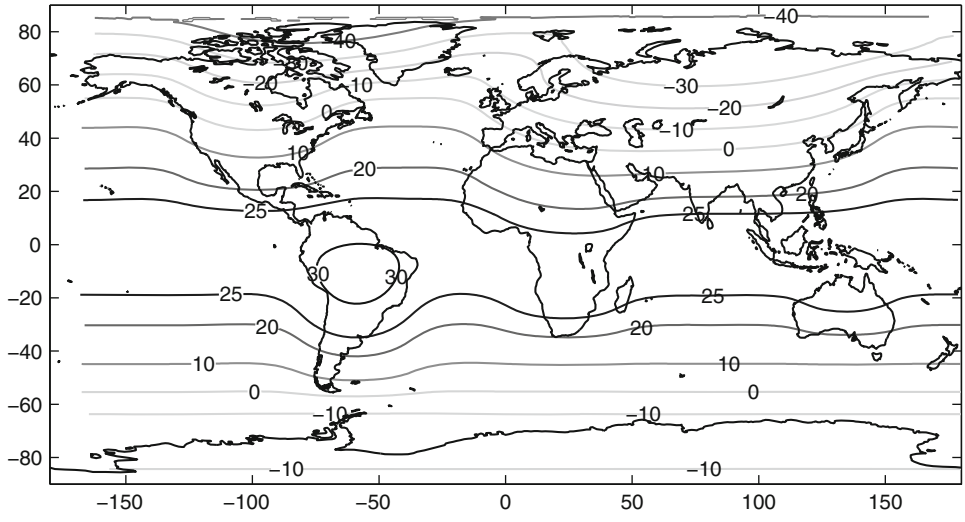


Figure 3
Distribution of January average temperature.

have presented the results for 10 years of simulations. Figure 3 shows the distribution of January temperature in the stationary periodic regime. Figure 4 displays the $-2°C$ snow lines for the Northern and Southern Hemispheres in January.

Figures 5 shows the distribution of temperature for the month of July, whereas the snow line for this month in both hemispheres is represented in Figure 6.

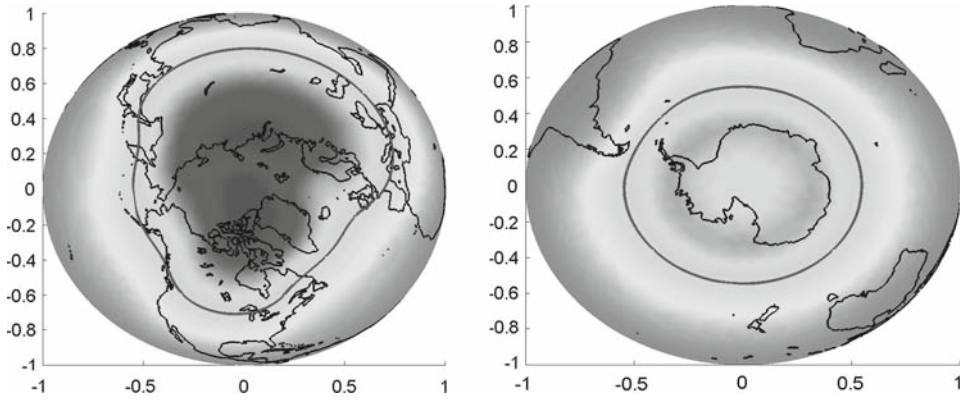


Figure 4
$-$ 2°C January snow line. Left: Northern Hemisphere; right: Southern Hemisphere.

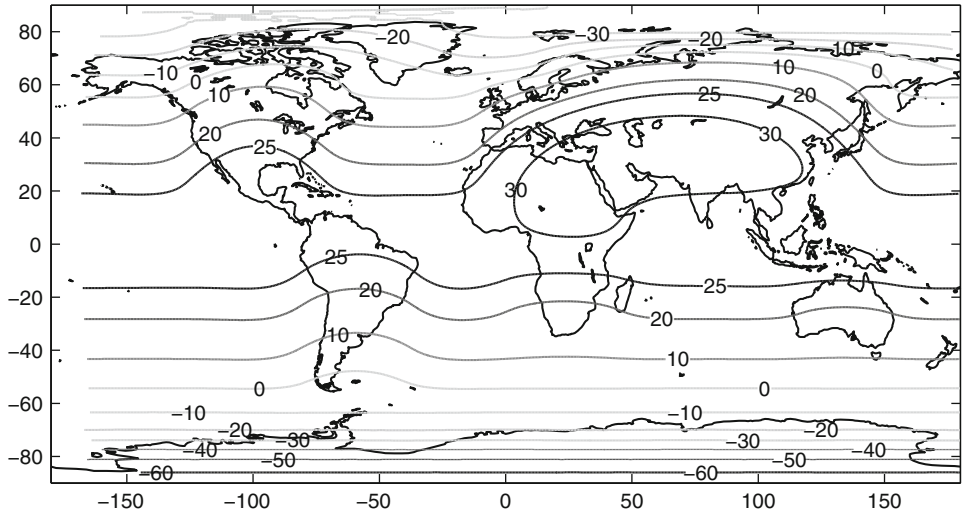


Figure 5
Distribution of average temperature for July.

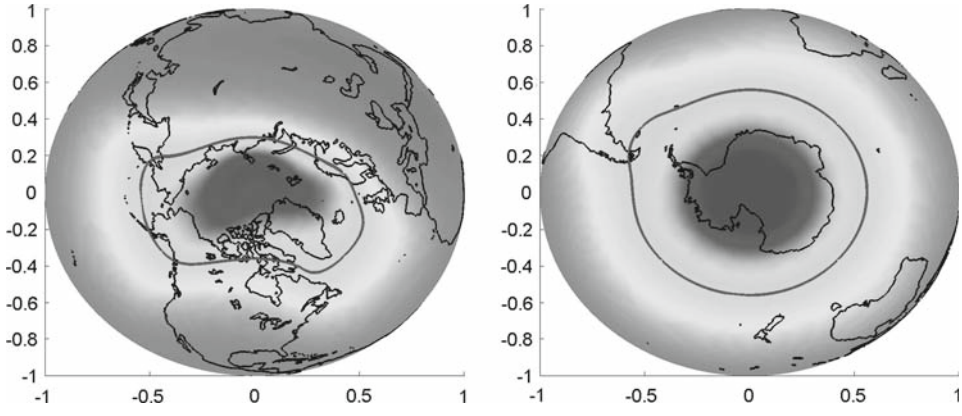


Figure 6
− 2°C July snow line. Left: Northern Hemisphere; right: Southern Hemisphere.

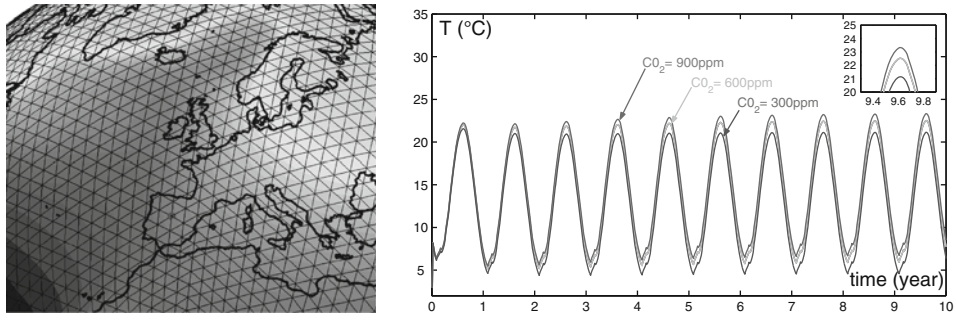


Figure 7
CO_2 influence on temperature at a point near Madrid. The box, showns the temperature corresponding to the month of July.

One can also simulate with our simple EBM the influence of CO_2 on the increase of temperature. We do so by considering that the concentration of CO_2 plays the role of an a additional forcing term $f(t,x)$ in the governing equation. Following Myhre *et al.* (1998) we model such a forcing as

$$f(t,x) = 5.35ln\left(\frac{C}{C_0}\right)\beta(x,u), \tag{15}$$

where $C_0 = 300$ ppm represents the concentration of CO_2 of preindustrial times and C is the value of concentration of CO_2 different of 300. Figure 7 displays the influence of the concentration of CO_2 on the temperature at a point near Madrid (Spain). We note that doubling the levels of CO_2 will produce an increase in the July and January average temperatures larger than 1.5°C.

Acknowledgments

The first author acknowledges the financial support from the Ministry of Education of Spain via research grant CGL2006-11264-C04-02. The second author is financially supported by a doctor scholarship from Consejería de Educación de Madrid y Fondo Social Europeo. The research of the third author is partially supported by the project MTM2005-03463 of the DGISGPI(Spain and CCG06-UCM/ESP-1110 of the DGUIC of the CAM and the UCM.

References

Bermejo, R., Carpio, J., Díaz, J. I, and Tello, L. (2007), *Mathematical and numerical analysis of a nonlinear diffusive climate energy balance model* (submitted).

Badii, M. and Díaz, J. I. (1999), *Time Periodic Solutions for a Diffusive Energy Balance Model in Climatology*, J. Math. Anal. Appl. *233*, 717–724.

Barrett, J. W. and Liu, W. B. (1994), *Finite element approximation of the parabolic p-Laplacian*, SIAM J. Numer. Anal. *31*, 413–428.

Baumgardner, J. R. and Frederickson, P. O. (1985), *Icosahedral discretization of the two-sphere*, SIAM J. Numer. Anal. *22*, 1107–1115.

Budyko, M. I. (1969), *The effects of solar radiation variations on the climate of the Earth*, Tellus *21*, 611–619.

Carl, S. (1992), *A combined variational-monotone iterative method for elliptic boundary value problems with discontinuous nonlinearities*, Applicable Analysis *43*, 21–45.

Díaz, J. I., *Mathematical analysis of some diffusive energy balance climate models*. In Mathematics, Climate and Environment (eds. Díaz, J. I., and Lions, J. L.) (Masson, Paris 1993) pp. 28–56.

Díaz, J. I. (Ed.) *The Mathematics of Models in Climatology and Environment*, ASI NATO Global Change Series I, no. *48* (Springer-Verlag, Heidelberg, 1996).

Díaz, J. I., Hernández, J., and Tello, L. (1997), *On the multiplicity of equilibrium solutions to a nonlinear diffusion equation on a manifold arising in Climatology*, J. Math. Anal. Appl. *216*, 593–613.

Díaz, J. I. and Hetzer, G. *A quasilinear functional reaction-diffusion equation arising in Climatology*, In *Equations aux derivees partielles et applications: Articles dedies a Jacques Louis Lions* (Gautier Villards, Paris 1998) pp. 461–480.

Díaz, J. I. and Tello, L. (1999), *A nonlinear parabolic problem on a Riemannian manifold without boundary arising in Climatology*, Collectanea Mathematica *50*, 19–51.

Dziuk, G., *Finite element for the Beltrami operator on arbitrary surfaces*. In *Partial Differential Equations and Calculus of Variations*, Lectures Notes in Mathematics, vol. 1357 (Springer, Heidelberg 1988) pp. 142–155.

Giraldo, F. X. and Warburton, T. (2005), *A nodal triangle-based spectral element method for the shallow water equations on the sphere*, J. Comput. Phys. *207*, 129–150.

Graves, C. E., Lee, W.-H. and North, G. R. (1993), *New parameterizations and sensitivities for simple climate models*, J. Geophys. Res. *98*, 5025–5036.

Hairer, E., Norsett, S. P. and Wanner, G., *Solving Ordinary Differential Equations I: Nonstiff Problems* (Springer-Verlag, Berlin, Heidelberg, 1993).

Heinze, T. and Hense, A. (2002), *The shallow water equations on the sphere and their Lagrange-Galerkin solution*, Meterol. Atmos. Phys. *81*, 129–137.

Hetzer, G. (1990), *The structure of the principal component for semilinear diffusion equations from energy balance climate models*, Houston Journal of Math. *16*, 203–216.

Hetzer, G., Jarausch, H. and Mackens, W. (1989), *A multiparameter sensitivity analysis of a 2D diffusive climate model*, Impact and Computing in Science and Engineering *1*, 327–393.

Hyde, W. T., Kim, K.-Y., Crowley, T. J. and North, G. R. (1990), *On the relation between polar continentality and climate: Studies with a nonlinear seasonal energy balance model*, J. Geophys. Res. *95* (D11), 18.653–18.668.

JU, N. (2000), *Numerical analysis of parabolic p-Laplacian. Approximation of trajectories*, SIAM J. Numer. Anal. *37*, 1861–1884.

MYHRE, G., HIGHWOOD, E. J., SHINE, K., and STORDAL, F. (1998), *New estimates of radiative forcing due to well mixed Greenhouse gases*, Geophys. Res. Lett. *25*, 2715–2718.

NORTH, G. R., *Multiple solutions in energy balance climate models*. In *Paleogeography, Paleoclimatology, Paleoecology 82* (Elsevier Science Publishers B.V., Amsterdam, 1990) pp. 225–235.

NORTH, G. R. and COAKLEY, J. A. (1979), *Differences between seasonal and mean annual energy balance model calculations of climate and climate sensitivity*, J. Atmos. Sci. *41*, 1189–1204.

SADOURNY, R., ARAKAWA, A. and MINTZ, Y. (1968), *Integration of the nondivergent barotropic vorticity equation with an icosahedral hexagonal grid for the sphere*, Mon. Wea. Rev. *96*, 351–356.

SELLERS, W. D., *Physical Climatology* (The University of Chicago Press, Chicago, Ill. 1965).

SELLERS, W. D. (1969), *A global climatic model based on the energy balance of the earth-atmosphere system*, J. Appl. Meteorol. *8*, 392–400.

WILLIAMSON, D. L. (1968), *Integration of the barotropic vorticity equation on a spherical geodesic grid*, Tellus *20*, 642–653.

XU, X. (1991), *Existence and regularity theorems for a free boundary problem governing a simple climate model*, Applicable Anal. *42*, 33–59.

(Received June 5, 2007, revised March 25, 2008, accepted April 5, 2008)
Published Online First: July 11, 2008

Pure appl. geophys. 165 (2008) 1049–1070
0033–4553/08/061049–22
DOI 10.1007/s00024-008-0346-4

Pure and Applied Geophysics

An Upper Limit to Ground Deformation in the Island of Tenerife, Canary Islands, for the Period 1997–2006

ANTONIO EFF-DARWICH,[1,2] OLIVIER GRASSIN,[2] and JOSÉ FERNÁNDEZ[3]

Abstract—Continuous monitoring of ground deformation in the volcanic island of Tenerife, Canary Islands, is based on GPS networks, since there are as yet no tiltmeter stations installed on the island. However, there is a world-class astronomical observatory on the island, the El Teide Observatory, where four tiltmeters, two aligned in the North-South and the other two in the East-West, are monitoring the movements of the solar telescope THEMIS. THEMIS (Heliographic Telescope for the Study of Solar Magnetism and Instabilites) is among the three largest solar telescopes in the world. Since THEMIS is located a few kilometers from the main volcanic structures of the island, in particular the El Teide-Pico Viejo stratovolcano, and the precision of the inclinometers is comparable to those used in geophysical studies, we carried out the analysis of the tilt measurements for the period 1997–2006. The tiltmeters at THEMIS are placed in the seventh floor of a tower, hence their sensitivity to geological processes is reduced compared to geophysical installations. However, THEMIS measurements are the only terrestrial data available in Tenerife for such a long period of observations, which include the sustained increase in seismic activity that started in 2001. In this sense, a significant change was found in the East-West tilt of approximately 35 μ-radians between the years 2000 and 2002. Some theoretical models were calculated and it was concluded that such tilt variation could not be due to dike intrusions, nor a volcanic reactivation below the El Teide-Pico Viejo volcano. The most likely explanation comes from dislocations produced by a secondary fault associated to a major submarine fault off the eastern coast of Tenerife. In any case, taking into account the nearly permanent data recording at THEMIS, they could be considered as a complement for any ground deformation monitoring system in the island.

Key words: Ground deformation, Tenerife, volcanic activity, earthquake.

1. Introduction

Some of the best astrophysical observatories in the world, namely the Canarian, Chilean and Hawaiian observatories, are located in active volcanic regions. This is not a coincidence, since topography modelled by volcanic activity is a main factor controlling the local atmospheric conditions and hence, the sky transparency that defines good astronomical sites.

[1] Departamento Edafología y Geología, Univ. de La Laguna, Facultad de Biología, Av. Astrofísico Francisco Sánchez s/n, 38206 Tenerife, Spain. E-mail: adarwich@ull.es

[2] THEMIS, S.L, 38200, La Laguna, Tenerife, Spain. Email: adarwich@themis.iac.es; grassin@themis.iac.es

[3] Instituto de Astronomía y Geodesia (CSIC-UCM), Facultad de Ciencias Matemáticas, Ciudad Universitaria, Plaza de Ciencias, 3, 28040 Madrid, Spain. Email: jose_fernandez@mat.ucm.es

The Canary Islands host two astrophysical observatories on two different active volcanic islands: Observatorio del Teide on the Island of Tenerife and Observatorio del Roque de los Muchachos in La Palma. Several volcanic eruptions have taken place on both islands, the last ones being in 1971 and 1909 for La Palma and Tenerife, respectively (CABRERA and HERNÁNDEZ-PACHECO, 1987).

Astrophysical observatories do not carry out geophysical monitoring programs, however there are secondary products from astrophysical observations that could be useful in geophysical studies. In particular, the pointing accuracy data collected by telescopes could contain information on crustal deformation that may complement data collected from geodetic techniques. Pointing accuracy should be in the order of the optical aberration introduced by the atmosphere in astronomical observations, namely lower than 2 arc-seconds or 9.7 μ-radians. Hence, spurious tilting motions of the telescope should be smaller than the optical aberration induced by the atmosphere. This tilting includes structural effects of the telescope, thermal deformation on the telescope building and geological activity, such as crustal deformation.

Geodetic techniques are being used extensively at active volcanoes and have provided useful eruption precursors (e.g., NEWHALL and DZURISIN, 1988; FISKE and SHEPHERD, 1990; DVORAK and DZURISIN, 1997; DZURISIN, 2007; FERNÁNDEZ *et al.*, 2005). Geodetic signals can play a key role in early detection of volcanic unrest due to the very high precision attainable with present-day techniques and instruments.

In this work, we present tilt data collected at the solar telescope THEMIS located at El Teide Observatory in the Canarian Island of Tenerife. We study the accuracy of these data and the limitations of their potential use in the framework of crustal deformation monitoring, complementing already existing geodetic networks and techniques.

2. Geological Setting

Tenerife is the largest island of the Canarian Archipelago and one of the largest volcanic islands in the world. It is located between latitudes 28–29° N and longitudes 16–17° W, 280 km distant from the African coast. It conforms an active volcanic region, its age varying from Middle Miocene to present, with no evidence of important gaps in its volcanic activity history, at least in the last 3 to 4 Ma (ANCOCHEA *et al.*, 1990, 1999). This activity is still evident in stationary low temperature fumarolic activity at Teide crater (< 85°C), diffusive gaseous emissions (VALENTÍN *et al.*, 1990; ALBERT-BELTRÁN *et al.*, 1990; HERNÁNDEZ *et al.*, 1998; PÉREZ *et al.*, 1996), groundwater temperatures reaching up to 50°C and volcanic contamination of groundwater in the subsurface of the central region (BRAVO *et al.*, 1976; CARRACEDO and SOLER, 1983; FARRUJIA *et al.*, 1994).

The morphology of Tenerife (see Fig. 1) is the result of a complex geological evolution: The subaerial part of the island was originally constructed by fissural eruptions of ankaramite, basanite and alkali basalts that occurred between 12 and 3.3 Ma (ANCOCHEA *et al.*, 1990; ARAÑA *et al.*, 2000; GUILLOU *et al.*, 2004). These formations

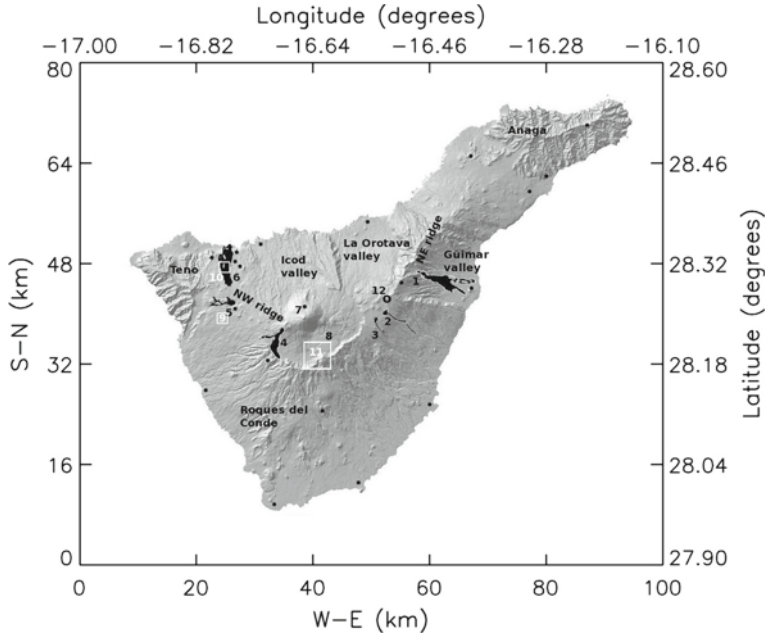


Figure 1

Simplified morphological map of Tenerife, including Teide-Pico Viejo stratovolcano (7), Las Cañadas Caldera (8) and the recorded historical eruptions (dark areas indicate extension of the eruption), namely Arafo (1), Fasnia (2), Siete Fuentes (3), Chahorra (4), Chinyero (5) and Arenas Negras (6). Black points indicate the approximate location of the GPS network, whereas the geodetic and levelling networks are enclosed by a white square (11). The subsidence areas detected by InSAR are represented by whites squares labelled as (9) for Garachico and (10) for Chio, whereas the location of the solar telescope THEMIS is represented by an open circle labelled as (12).

made up shield volcanoes that remain at present as three eroded massifs occupying the three corners of the island (Teno, Anaga and Roques del Conde massifs). In the central part of the island, from 3.5 Ma to present, the emission of basalts and differentiated volcanics gave rise to a large central volcanic complex, the Las Cañadas Edifice (MARTÍ *et al*, 1994). After a period of mafic volcanism, several periods of phonolitic activity took place, culminating in the formation of a large elliptical depression measuring 16×9 km^2, known as Las Cañadas Caldera. In the northern sector of the caldera, the Teide-Pico Viejo complex was constructed as the product of the most recent phase of central volcanism. Teide-Pico Viejo is a large stratovolcano that has grown during the last 175 Ky. The post-shield basaltic activity, which overlaps the Las Cañadas Edifice, is mainly found on two ridges (NE and NW), which converge on the central part of the island (ANCOCHEA *et al*., 1990; CARRACEDO, 1994; ABLAY and HURLIMANN, 2000). Large-scale lateral collapses, involving rapid mass movements of hundreds of cubic kilometers of rock, are responsible for the formation of three valleys: La Orotava, Güimar and Icod. Recorded eruptive activity has consisted of six Strombolian eruptions (CABRERA and

HERNÁNDEZ-PACHECO, 1987), namely Siete Fuentes (1704), Fasnia (1705), Arafo (1705), Arenas Negras (1706), Chahorra (1798) and Chinyero (1909). The last three eruptions occurred at the NW axis system, the most active area of the island together with El Teide-Pico Viejo Edifice for the last 50,000 years (CARRACEDO *et al.*, 2003a, b).

It is important to mention, in the context of geodetic monitoring, the possible existence of a shallow magma chamber underneath Teide-Pico Viejo. It is estimated from petrologic analyses (ARAÑA 1985; ARAÑA *et al.*, 1989) that the top of the magma chamber is located at sea level, having a volume of approximately 30 km^3 and a radius of 2 km (under the supposition of spherical shape). Thermodynamical (DIEZ and ALBERT, 1989) and chemical (ALBERT-BELTRÁN *et al.*, 1990) modelling of the fumaroles at El Teide summit revealed that the present temperature at the surface of the magma chamber would be approximately 350°C, whereas the top of the chamber coincides with that calculated from petrologic analyses. However, ARAÑA *et al.* (2000) found long wavelength magnetic anomalies in the central part of Tenerife that could be interpreted as the top of deep intrusive bodies or magma chambers zone ($\approx$5.7 km b.s.l.). In this sense, the possible location of the top of the magma chamber ranges from sea level to nearly 6 kilometers below sea level.

It is also important for geodetic studies that most recent eruptions (<3 Ma) have been fed by dikes (FERNÁNDEZ *et al.*, 2003). These dikes are associated to systems of deep fractures that generally respond to regional tectonics. In other cases, the dikes are located in shallow radial or circular fractures in large volcanic structures. Most of the visible dikes are less than one meter thick in the shallowest sections. However, when erosion exposes deeper sections, they are seen to be much thicker, especially those of a saline composition. Regional fractures in Tenerife are mainly to be seen in the two ridges (NE and NW) that converge in the central region of the island. There are also major radial fractures associated to the eruptive systems of the Teide-Pico Viejo volcano, in the central area of Tenerife.

3. Previous Geodetic Studies in the Island of Tenerife

In recent decades several observational and theoretical studies have been carried out in the context of geodetic monitoring in the volcanic Island of Tenerife. A 17-benchmark classical geodetic network and a levelling profile located in the area of Las Cañadas Caldera (see Fig. 1) have been observed several times since 1982 (SEVILLA and MARTÍN, 1986; SEVILLA *et al.*, 1996). No displacements were found from 1982 to 2000 during the observation of both the geodetic network and the levelling profile (FERNÁNDEZ *et al.*, 2003, 2005). Theoretical analysis carried out by YU *et al.* (2000) demonstrated the need to extend the existing geodetic network in Las Cañadas Caldera to cover the full island for volcano monitoring purposes.

Ground displacement analyses have been carried out on the entire island by means of classical DInSAR techniques for the period 1992–2000 (CARRASCO *et al.*, 2000; FERNÁNDEZ *et al.*, 2002, 2005; ROMERO *et al.*, 2002). These works did not reveal any

significant ground deformation on Las Cañadas Caldera; however, two deformation areas were detected for the period July 1993 to September 2000 (see Fig. 1), extending over 15 km^2 with a ground subsidence of 10 cm (Garachico deformation) and 8 km^2 with a ground subsidence of 3 cm (Chio deformation). These subsidences were found closer to the locations of the most recent eruptions in the island (Arenas Negras, Chahorra and Chinyero).

Considering the results by YU *et al*. (2000) and to validate DInSAR results, a 22-station GPS network was designed and observed in August 2000, with more densification in the two deformation areas (FERNÁNDEZ *et al*., 2003). The densified network was re-observed in 2001 and 2002 (FERNÁNDEZ *et al*., 2004), confirming both subsidences previously detected by DInSAR. Examination of the geophysical observations on the island, human activities underway and the results of the theoretical modelling seem to indicate that at least part of the observed deformation may be caused by changes in the groundwater level and therefore that part of the deformation should not be linked to a volcanic reactivation. This result is important because it implies that, if geodetic volcano monitoring is to be performed on the island, the system used must be capable of discerning between various possible origins of the deformation by analyzing their patterns and ancillary information from other sources. Another important result obtained by FERNÁNDEZ *et al*. (2004) is a change in the deformation pattern in the period 2000–2002 with respect to that observed by DInSAR for the period 1992–2000. This could relate to the change in seismicity detected in the same period and could be a geodetic precursor of the 2004 volcano-tectonic crisis (FERNÁNDEZ *et al*., 2006a).

4. Instrumental Setting and Data Acquisition

THEMIS (Heliographic Telescope for the Study of Solar Magnetism and Instabilites) is among the three largest solar telescopes in the world. It is devoted to the analysis of magnetism and the dynamics of the solar atmosphere. THEMIS belongs to a French-Italian consortium (CNRS-CNR) and it is located at the Observatorio del Teide (Tenerife, Canary Islands, Longitude: 16°30′35″ W, Latitude: 28°18′00″ N, Height: 2400 m.a.s.l.), a world-class astrophysical observatory managed by the Instituto de Astrofísica de Canarias. The observatory lies within 5 kilometers of Las Cañadas Caldera wall and it is just 12 kilometers distant from El Teide-Pico Viejo volcano summit, 20 kilometers distant from the NW ridge and within a few kilometers of La Orotava valley, Güimar valley and the NE ridge. Hence, THEMIS and in general the observatory, is revealed as an ideal location to monitor the main active volcanic areas of Tenerife.

As was mentioned in the previous section, pointing accuracy is one of the main concerns when operating a telescope. THEMIS, as with many other solar telescopes, consists of a tower on top of which the telescope is placed. Such design is intented to reduce atmospheric turbulence induced by soil heating. However, the building could tilt due to telescope operations, wind, temperature changes and/or geological variables such as earthquakes and ground deformations. In an attempt to reduce meteorological tilting

effects, the telescope is divided into two mechanically-isolated cylindrical buildings (see Fig. 2), namely an internal pillar where the telescope is placed and the external building where the dome, laboratories and offices are located. The external building will also absorb the effects of wind and large atmospheric temperature changes, whereas the internal pillar will be mainly affected by telescope operations; geological variables will affect both the internal and external buildings. In any case, both the pillar and the external building are equipped with two North-South and East-West tiltmeters to continuously monitor variations in the inclination of the building and hence, the pointing accuracy of the telescope.

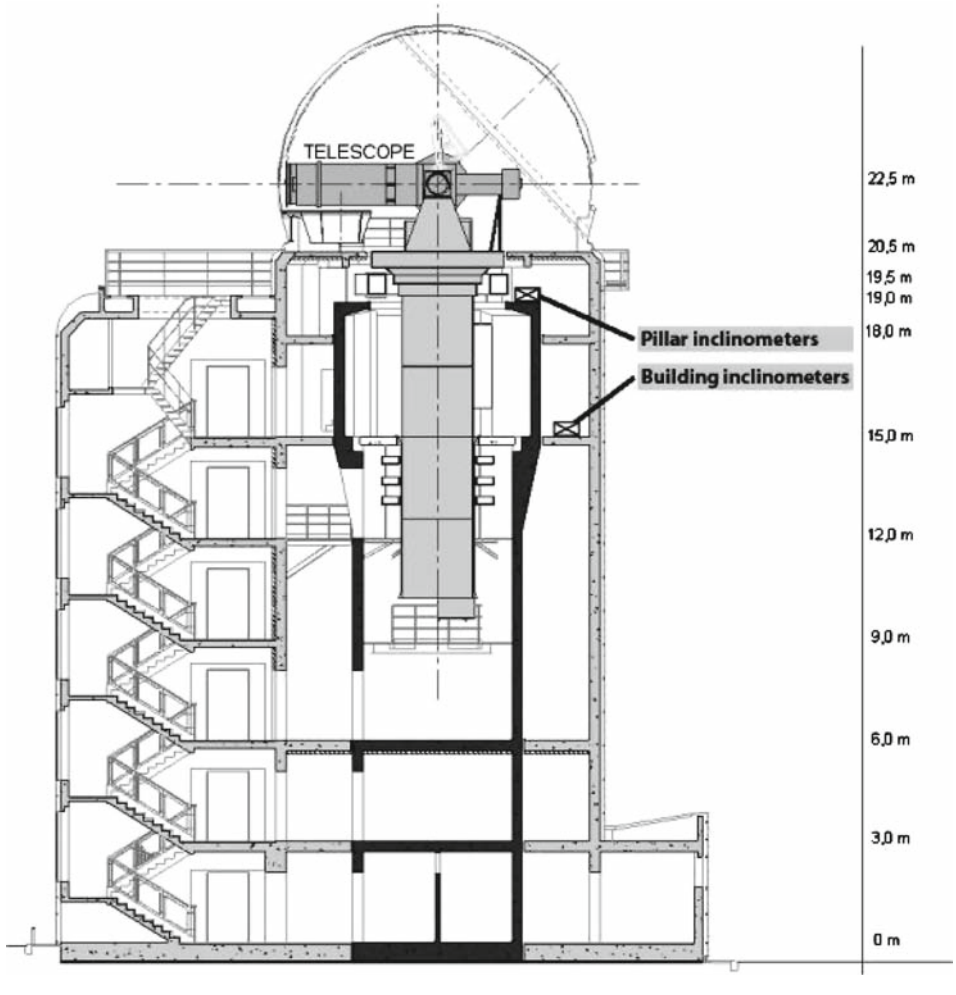


Figure 2
Sketch of the internal structure of THEMIS. Dark shaded areas represent the internal pillar, whereas light-grey shaded areas correspond to the external building. The location of the inclinometers is also shown.

The tiltmeters were designed and built at the Institute de Physique du Globe of Paris, being based on the Blum pendulum (SALEH *et al*., 1991) and entirely built out of monolithic silica to minimize temperature deformation. An opto-electronic sensor provides a signal proportional to illumination, according to the pendulum position relative to a vertical pseudo-axis. The tiltmeters are encased to protect them against light and electrostatic effects. Both the pillar and the external building tiltmeters are grouped in so-called tiltmeter units. The two units also provide continuous measurement of temperature and conditioning of the analogical signal to filter out micro-seismic components and other interferences.

The analogical signal provided by the inclinometer units is converted into digital signal at a sampling rate of 1 measurement per minute. The accuracy of the tilt signal is approximately 0.02 μ-radians, whereas the accuracy for temperature measurements is 0.01°C.

5. *Data Analysis*

In the present work, we have only considered the data obtained from the pillar tiltmeters, since the external building measurements are strongly affected by wind and atmospheric temperature changes. Telescope observations also affect the tilt signal and hence, we only used data recorded from midnight to 6:00 AM, corresponding to the temporal interval in which the telescope is not observing and hence, it is not moving. The original sampling rate of one measurement per minute was hence reduced to one averaged measurement per day. Time series for daily averages of the tilt and temperature signals for the entire observing range (1 January, 1997 to 20 May, 2006), as well as their spectral response, are shown in Figures 3 and 4, respectively. Long-term (several days to yearly) variations in the tilt signals are clearly modulated by temperature. This modulation is responsible for the increment in the spectral amplitude of the tilt signals with the period, as illustrated in Figure 4. Tilt measurements are clearly modulated by temperature and hence, it is necessary to filter out this effect, in order to study the possible influence of other variables, such as crustal deformation. However, at short periods, the effect of temperature is negligible and the spectral amplitude of the tilt signal lies below 1 μ-radian (see Fig. 4). Hence, it is possible to carry out high accuracy measurements of short-period variations in the tilt signal, like those expected by a distant earthquake or a dyke injection (IRWAN *et al*., 2003).

Since long-term temporal variations in both tilt and temperature signals are similar, we will perform a temperature filtering process from the tilt signal in the frequency domain using wavelet analysis. A practical introduction and description to wavelet analysis can be found e.g., in TORRENCE and COMPO (1998) and MEYERS *et al*. (1993). Wavelets are mathematical functions, that in our case consist of wave packets, $M(n, s)$, defined as a plane wave of a given frequency modulated by a Gaussian function of a given width (Morlet function). The correlation between the wavelet and our data (a time

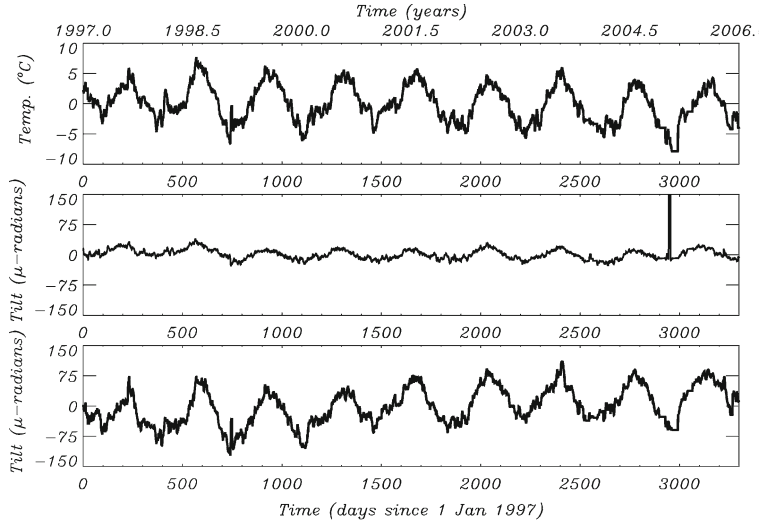


Figure 3

Temperature (upper panel), North-South tilt (central panel) and East-West tilt (lower panel) time-series recorded by the pillar inclinometer unit for the period January 1, 1997 to January 1, 2006. Increasing North-South tilt with time represents downward deformation to the North, whereas increasing East-West tilt with time represents downward deformation to the East.

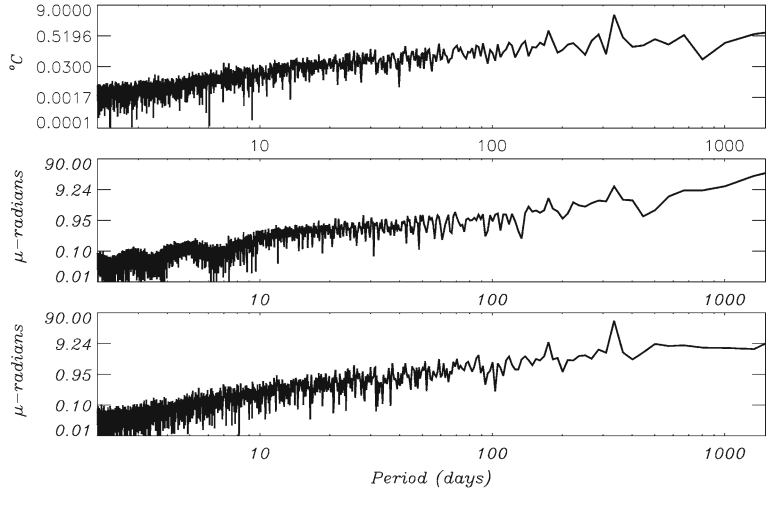


Figure 4

Temperature (upper panel), North-South tilt (central panel) and East-West tilt (lower panel) spectra of the time series collected at THEMIS for the period 1997–2006.

series X, with values of x_n at time index n) gives a local measure of the projection of the wave packet in our data, in other words it gives us a local measure of the relative amplitude of activity at a given frequency of the wave packet. By sliding this wavelet

(with translation parameter n) along our time series, and varying the "scale" of the wavelet by changing its width (with dilation parameter s) one can then construct a new time series $W_n(s)$ of the projection amplitude versus time,

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'} M^*(n', s) \frac{(n' - n)dt}{s}, \tag{1}$$

where the asterisk denotes complex conjugate and dt is the constant time interval between two values x_n. From this definition, it is also possible to obtain x_n at a given time step n, from a linear combination of the wavelet functions $W_n(s)$.

$$x_n = \sum_{s'=0}^{S-1} C(n, s') W_n(s'), \tag{2}$$

where $C(n,s')$ are the coefficients for the linear combination (see TORRENCE and COMPO 1998 for details). The advantage of this representation for a series X is that we could filter out unwanted oscillations (with period s_u) in the signal by excluding in the linear combination (2) those coefficients associated to the dilation parameter $s = s_u$.

Let us perform wavelet analysis on the normalized time series for temperature P(n), North-South tilt $N(n)$ and East-West tilt $E(n)$, as illustrated in Figure 5. In this figure, it is clearly present the yearly (for periods of approximately 365 days) modulation of both temperature and tilt signals. At each time step n the ratios $R_N(s_j)$ and $R_E(s_j)$ are calculated:

$$R_N(s_j) = 1 - abs\left(\frac{C_N(n, s_j) W_n(s_j) - C_p(n, s_j) W_n(s_j)}{C_N(n, s_j) W_n(s_j) + C_p(n, s_j) W_n(s_j)}\right), \tag{3}$$

$$R_E(s_j) = 1 - abs\left(\frac{C_E(n, s_j) W_n(s_j) - C_p(n, s_j) W_n(s_j)}{C_E(n, s_j) W_n(s_j) + C_p(n, s_j) W_n(s_j)}\right), \tag{4}$$

where $C_P(n, s_j)$, $C_N(n, s_j)$ and $C_E(n, s_j)$ are the linear coefficients defined in equation (2) for temperature, North-South tilt and East-West tilt, respectively. If $R_N(s_j)$ or $R_E(s_j)$ are larger than 0.95, the associated coefficient $C_j(n, s_j)$ in equation (2) are excluded from the linear combinations to recover $N(n)$ and $E(n)$ from the corresponding wavelet functions. In this way, we filtered out the temperature temporal modulation from the tilt signals, as shown in Figure 6, 15 μ-radians being the minimum level of detection for the filtered signals.

6. Analysis of the Results

The origin of the temperature-filtered tilt signal is difficult to assess, recalling that the tilt sensors are placed in the seventh floor of a building. In this sense, temperature dilation

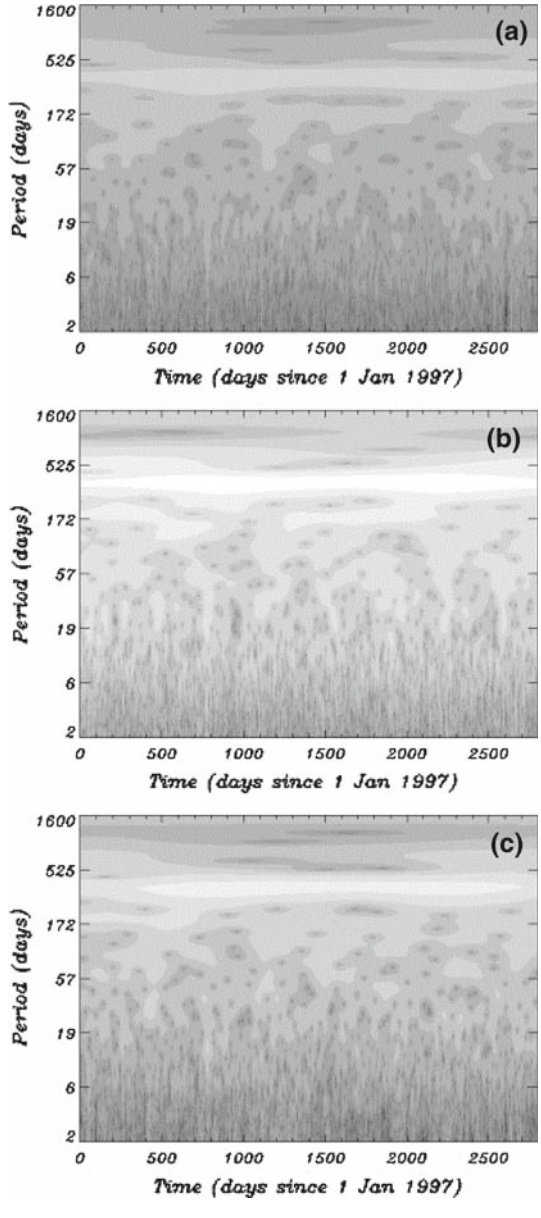


Figure 5

(a) Wavelet analysis for the time series associated to the North-South tilt. **(b)** Wavelet analysis for the time series associated to the East-West tilt. **(c)** Wavelet analysis for the time series associated to the temperature.

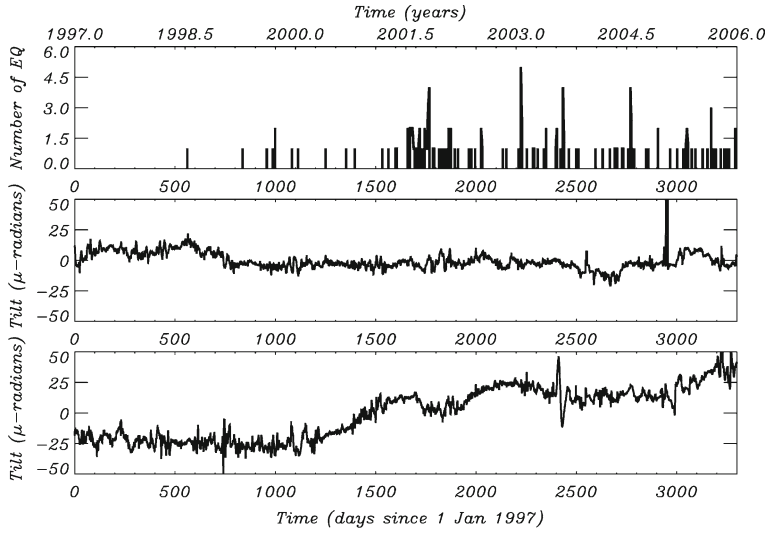


Figure 6
Residual time series for the North-South (central panel) and East-West (lower panel) tilt measurements after filtering out temperature effects through wavelet analysis. The upper panel shows the number of earthquakes with magnitude larger that 2.3 detected in and around Tenerife.

effects on the pillar and/or motion transmission from the external building through the subsurface should not be disregarded.

In an attempt to relate the residual tilt signal to geological variables, the temporal and spatial distributions of earthquakes detected by the Spanish Instituto Geográfico Nacional (IGN) in and around Tenerife were calculated, as shown in Figures 7–9. Only earthquakes with magnitudes equal to or larger than 2.3 (Richter scale) were considered, since information on lower magnitude earthquakes could be inaccurate due to reconfigurations in the local seismic network (M.J. Blanco from IGN, private communication). The maximum magnitude recorded during the period 1997–2006 was 3.4, whereas the largest instrumental earthquake in this area was recorded on 9 May 1989, reaching a magnitude of 5.2. The focal mechanism of this earthquake shows strike-slip movements with two nodal planes oriented north-northeast-south-southwest and north-west-southwest. The former agrees with the aftershock distribution and the strike of a major submarine fault parallel to the eastern coast of Tenerife (BOSSHARD and MACFARLANE, 1970; MEZCUA *et al.,* 1990; GONZÁLEZ DE VALLEJO *et al*., 2006). Moreover, a series of liquefaction-related structures (namely clastic dykes and tubular vents) were discovered in Holocene sand deposits in southern Tenerife (GONZALEZ DE VALLEJO *et al*., 2003), likely being the result of an estimated 6.8 magnitude earthquake in the submarine fault. Starting in 2001, there is a significant increase in seismic activity that is primarily taking place at the location of the submarine fault. However, part of the seismic activity moved inland, following a SE-NW trend, as illustrated in Figure 8. During 2004, the

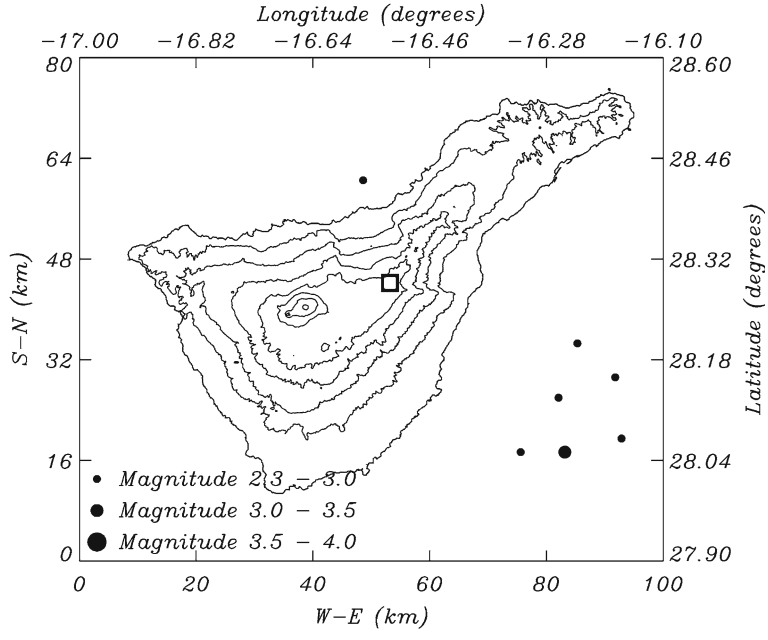


Figure 7
Spatial distribution of the earthquakes located in and around Tenerife for the period 1997–1999. Radius of the circles is proportional to the magnitude of the earthquake. Only earthquakes with a magnitude larger or equal to 2.3 are represented. Black square shows the position of the Observatorio del Teide.

increase in seismic activity continued, however part of the inland activity moved westwards, to an area with no previous instrumentally recorded seismic activity (see Fig. 9). Unlike the 1989 earthquake, there is no accurate information regarding the location of the hypocenters and the focal mechanisms of the earthquakes that occurred between 1997 and 2006.

A significant increment in diffuse gas emissions was also recorded in the NW ridge during 2004, after several soil-gas sampling campaigns (PÉREZ *et al.*, 2005). In this work, however, we will study the sustained increase in seismic activity that started in 2001, since this was proceeded and accompanied by a significant variation in the East-West tilt signal of approximately 35 μ-radians (between days 1000 and 2000 in Fig. 6).

We carried out a theoretical analysis in order to study the possible relation between tilt variation and volcano-tectonic processes, namely activity associated to a magma chamber, a dike injection and a dislocation induced by a fault. The first case we studied was the sensitivity of the THEMIS tiltmeters (hereafter TT) measurements to reactivation in the magmatic system associated to El Teide-Pico Viejo stratovolcano. We considered the crustal structure model described in Table 1 and the elastic-gravitational deformation model described by RUNDLE (1982), FERNÁNDEZ and RUNDLE (1994) and FERNÁNDEZ *et al.*

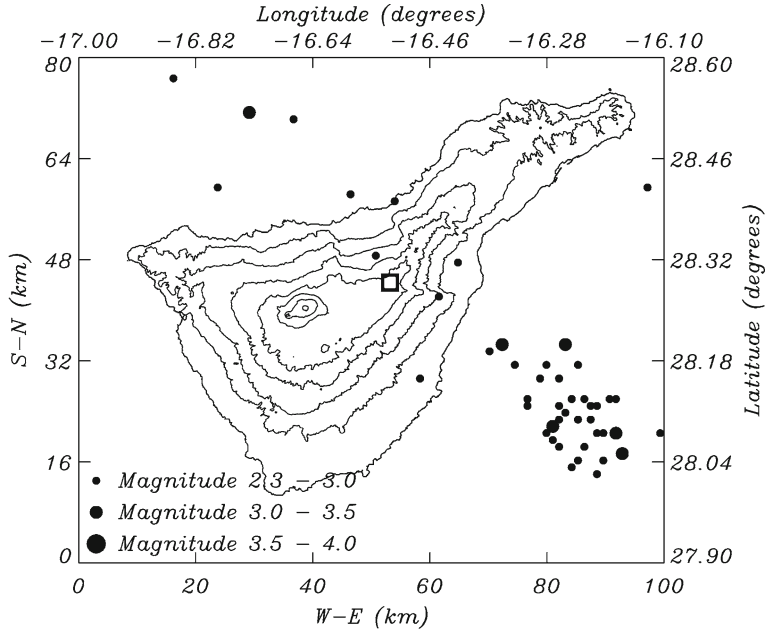


Figure 8
As in Figure 7, but for the period 2000–2002.

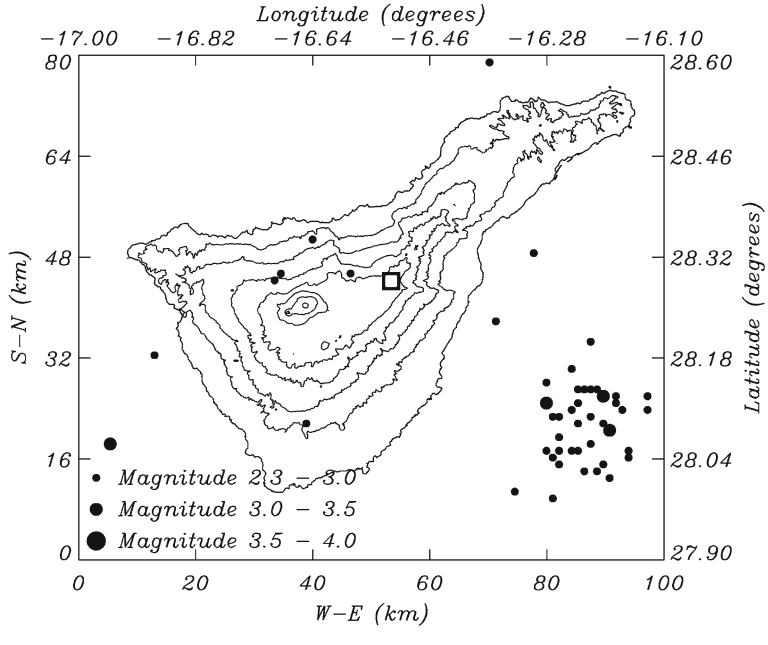


Figure 9
As in Figure 7, but for the period 2003–2006.

Table 1

Tenerife crustal model (FERNÁNDEZ et al., 1999)

Layer	Thickness (km)	ρ (10^3 kg m^{-3})	μ (10^{10} Pa)	λ (10^{10} Pa)
1	3.5	2.1	0.7	0.8
2	2.5	2.3	1.3	1.7
3	19.	2.9	4.0	4.7
Mantle		3.3	6.4	8.3

(1997, 2006b). As was mentioned, the possible location of the magma chamber underneath El Teide-Pico Viejo is uncertain, ranging from the position of the top of the chamber from approximately sea level to 6 kilometers below sea level. In this sense, several calculations were carried out for different cases of spherical intrusion, varying its depth and the parameter Pa^3, P and a being the pressure and radius of the magma intrusion (FERNÁNDEZ *et al.*, 1997), respectively. Figures 10 and 11 present absolute ground displacement and tilt for intrusions located underneath El Teide-Pico Viejo stratovolcano at 2, 6 and 10 kilometers depth, that are characterized by two different values of Pa^3, *namely* 10^3 and 10^4 MPa/km^3. TT is located approximately 12 kilometres

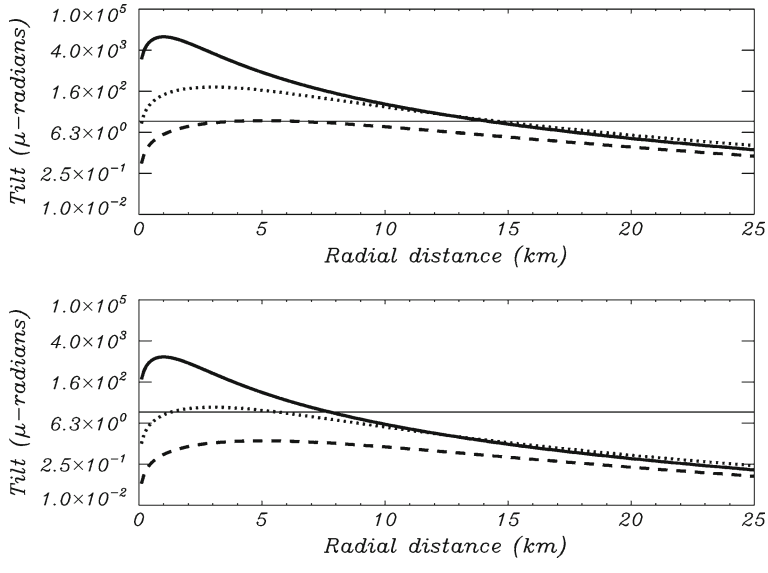


Figure 10

Absolute value for the displacement (in millimeters) as a function of distance from the projection onto the surface of different states of spherical magma intrusions (radius, depth and pressure). Upper and lower panels present results for instrusion with Pa^3 = 10^4 MPa/km^3 and Pa^3 = 10^3 MPa/km^3, respectively (see text for details). Solid, dotted and dashed lines correspond to the results obtained for magma intrusions located at a depth of 2, 6 and 10 kilometers below sea level.

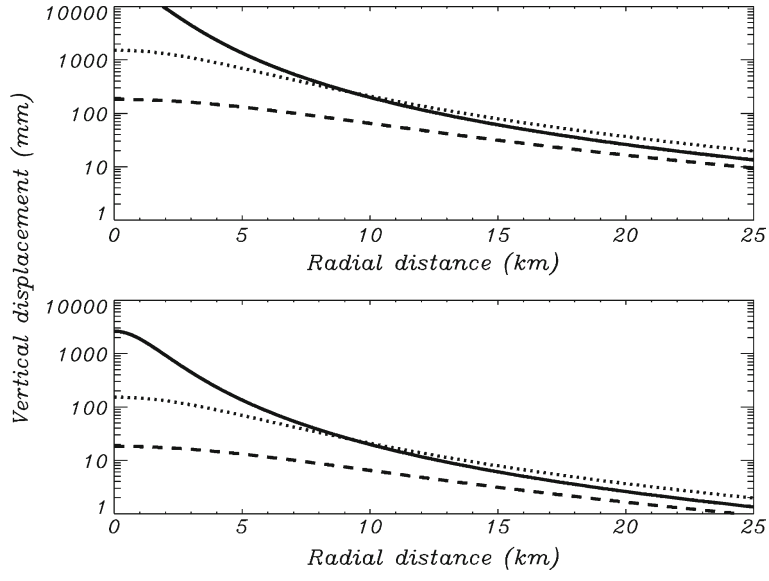


Figure 11

Absolute value for the tilt as a function of distance from the projection onto the surface of different states of spherical magma intrusions (radius, depth and pressure). Upper and lower panels present results for intrusions with $Pa^3 = 10^4$ MPa/km^3 and $Pa^3 = 10^3$ MPa/km^3, respectively (see text for details). Solid, dotted and dashed lines correspond to the results obtained for magma intrusions located at a depth of 2, 6 and 10 kilometers below sea level. Thin continuous line indicates the sensitivity of the THEMIS tiltmeters.

eastwards of El Teide-Pico Viejo and hence, it is sensitive to some cases of magmatic intrusions, namely at depths of up to 6 kilometers below sea level and $Pa^3 = 10^4$ MPa/km^3. However, the detectable cases imply vertical displacements larger than 10 cm at TT and larger than 1 meter at the summit of El Teide-Pico Viejo that have not been detected with GPS or DInSAR observations during the studied time period (FERNÁNDEZ *et al.*, 2004, 2005). Moreover, it is expected that the North-South and East-West tilts were similar, since we assume spherical symmetry. However, in the case presented in Figure 6, the East-West tilt is significantly larger and hence, it is not likely that a magma intrusion below El Teide-Pico Viejo volcano could induce the tilt measured by TT for the period 2000–2002.

The second case of the theoretical analysis corresponds to the other possible kind of magmatic intrusion in Tenerife, namely dikes (MARINONI and GUDMUNDSSON, 2000; YU *et al.*, 2000). To explain the observed ground deformation, we use the conventional assumption of dislocations buried in an elastic half space composed of a Poisson solid (OKADA, 1985). We specified eight parameters describing the rectangular fault patch: Three centroid coordinates X, Y, h, strike angle α and dip angle δ along-strike length L and down-dip width W, and the slip vector of the tensile component (thickness of the dike) U_3, as illustrated by FEIGL and DUPRÉ (1999). These authors developed the

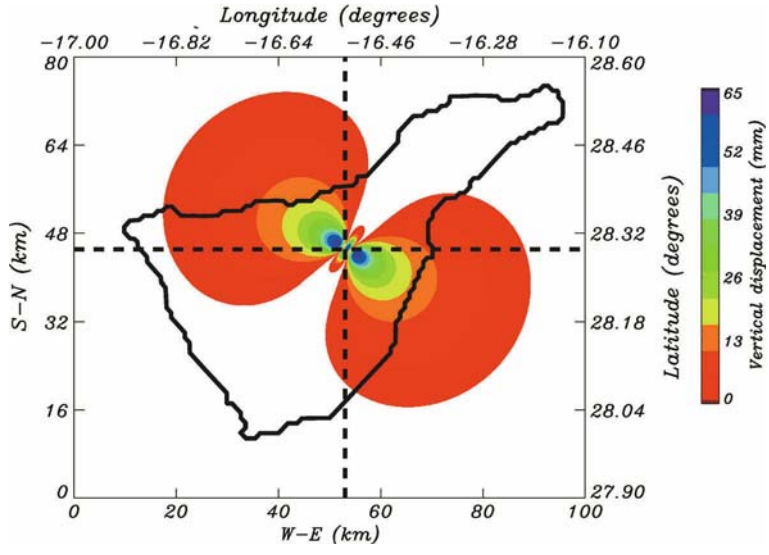


Figure 12
Absolute value for the vertical displacement (in millimeters) induced by a dike intrusion in the NE ridge of Tenerife. Vertical and horizontal dashed lines cross at the position of the Observatorio del Teide.

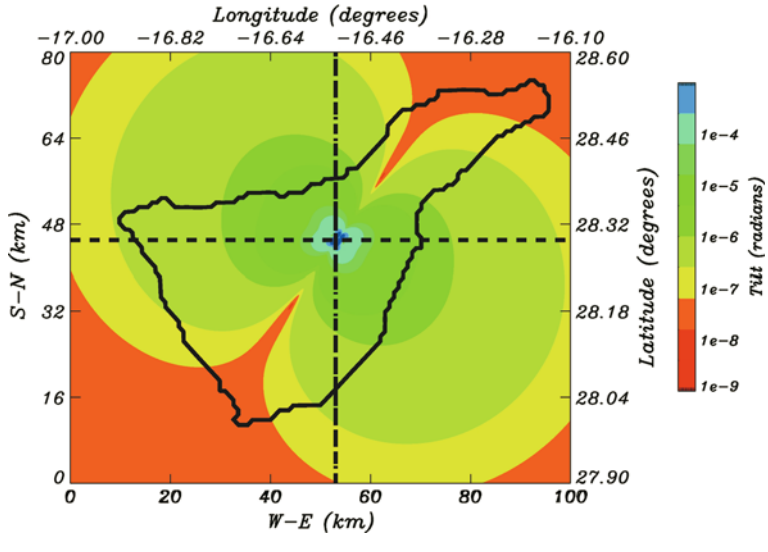


Figure 13
Absolute value for the tilt induced by the dike intrusion illustrated in Figure 12. Vertical and horizontal dashed lines cross at the position of the Observatorio del Teide.

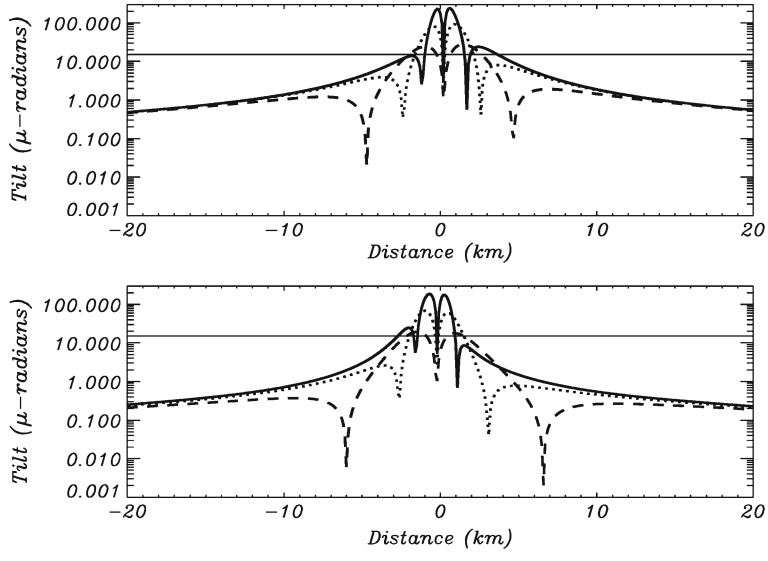


Figure 14
Tilts calculated along the vertical (upper panel) and horizontal (lower panel) dashed lines shown in Figure 13. Solid, dotted and dashed lines represent the tilts calculated for a dike intrusion located 500 meters, 1 kilometer and 2 kilometers from the surface, respectively.

numerical code RNGCHN that was used to calculate the ground deformation due to dike intrusion. Figures 12 and 13 show the effects produced by a dike located in the NE ridge, extending from the mantle (about 25 km depth) to 1 km depth, $2L = 2$ km, $U_3 = 1$ m, and $\delta = 90°$. Figure 14 presents the tilt along the dashed lines shown in Figures 12 and 13 for the cases of a dike extending from the mantle to 2 km, 1 km and 0.5 km, respectively. The dike is intruding perpendicularly to ground surface and hence, the amplitude of the displacement is proportional to U_3, in the sense that a dike 2 meters thick will induce ground deformations twice those presented for a 1 meter thick dike in Figures 12 to 14. Since the sensitivity of TT is approximately 15 μ-radians, only shallow dikes (above 2 km in depth) located in the vicinity of the telescope (at distances not exceeding 5 km) could be detected, even for dikes several meters thick. However, tilt variations are not accompanied by an increase in seismic activity (there are no epicenters in the area), as it could be expected during a dike intrusion (e.g., YAMAOKA *et al.*, 2005). Moreover, the duration of the change in the East-West tilt spans more than two years, being an excessively long period to be caused by a dike intrusion. Hence, it is unlikely the tilts in Figure 6 are due to a dike; however, an interesting result could be extracted from Figures 12 to 14: A high precision tiltmeter (in the order of nano-radians) is nearly sensitive to a dike intruding anywhere within the island. In this sense, a network of at least 8 high precision tiltmeters

deployed throughout the island could be sufficient to infer the 8 parameters that characterize a dike, namely coordinates *X, Y, h*; angles α and δ along-strike length *L* down-dip, width *W* and U_3.

The third case in our theoretical analysis is represented by a dislocation induced by a fault. Following FEIGL and DUPRÉ (1999), we specified nine parameters describing the rectangular fault patch: Three centroid coordinates *X, Y, h*; strike α and dip δ; along-strike length *L* and down-dip width *W*; and the slip vector of the left-lateral U_1 and up-dip U_2 components. Following the inland lineation of earthquakes for the period 2000 and 2002, that runs parallel to the existing volcanic seamount alignment described by ROMERO-RUIZ *et al*. (2000), we surmised that the tilt variation detected at THEMIS could be the result of a nearly strike-slip secondary fault associated to the major submarine fault parallel to the eastern coast of Tenerife (Figures 15 and 16). As in the case of the dike intrusion, it is not possible to find a unique solution, since there are nine unknown parameters and only two observables, namely the North-South and the East-West tilts. The results presented in Figures 15 and 16 correspond to a fault extending from 5 km below the surface to a depth of 10 km with 5 cm of left lateral strike-slip and 10 cm of dip-slip. This is a simple solution about the effect of a fault that could explain both the tilt measured at THEMIS and the seismic activity that took place along the secondary fault line.

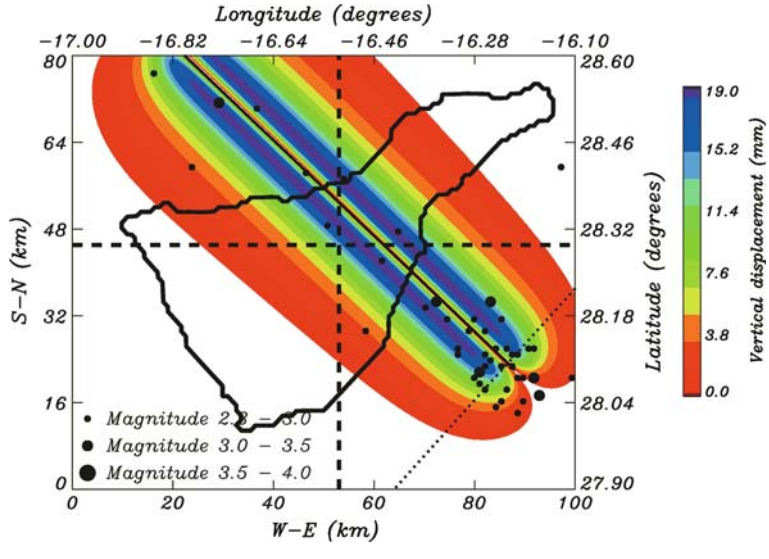


Figure 15

Absolute value for the vertical displacement (in millimeters) induced by a strike-slip fault perpendicular to the main submarine fault off the coast of Tenerife. Thick dotted and solid lines represent the projection on the surface of the major submarine fault and the secondary fault (see text for details). Filled circles represent the location of the earthquakes detected between 2000 and 2002. Vertical and horizontal dashed lines cross at the position of the Observatorio del Teide.

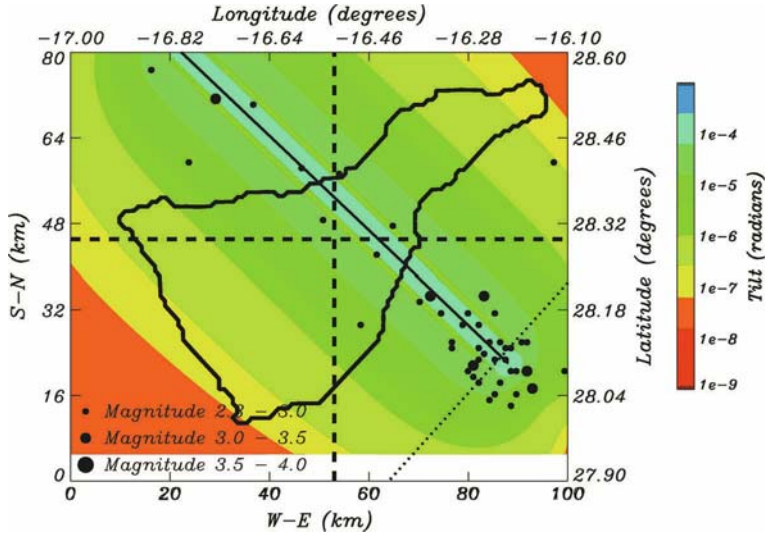


Figure 16
Absolute value for the tilt calculated from the vertical displacement calculated in Figure 15. Thick dotted and solid lines represent the projection on the surface of the major submarine fault and the secondary fault (see text for details). Filled circles represent the location of the earthquakes detected between 2000 and 2002. Vertical and horizontal dashed lines cross at the position of the Observatorio del Teide.

7. Conclusions

The methodology used in the acquisition and processing of tiltmeter data at THEMIS sets the minima detectable tilts to 15 μ-radians. This limits the applicability of THEMIS data for volcano monitoring in the island of Tenerife. In any case, a significant change was found in the East-West tilt of approximately 35 μ-radians between the years 2000 and 2002. From theoretical study, and recalling that all theoretical results are based on simple homogeneous elastic half-space models that do not take into account topographic effects or lateral heterogeneities, we can conclude that tilt produced by dike intrusions could only be detected in cases of great shallow dikes in the NE ridge less than 5 km distant from TT. The effects of dike intrusions in the NW ridge could not be detected with the tiltmeters at THEMIS. If we consider a volcanic reactivation below the El Teide-Pico Viejo volcano, only great ($Pa^3 >> 10^3$ MPa/km^3) and shallow events (center of mass located at 6 to 2 km depth) could produce effects measurable by these instruments. In the case of tectonic processes, THEMIS tiltmeters could detect some cases of dislocation produced by the faults associated to the major submarine fault located off the eastern coast of Tenerife.

In summary, considering the limitations in the applicability of THEMIS data for geodetic monitoring and considering their nearly permanent data recording for the last ten years, tiltmeter monitoring at THEMIS could be considered complementary to any monitoring system in the island, since they could detect (at least theoretically)

some cases of activity, mainly slow tectonic processes and massive magmatic intrusions.

Acknowledgements

AED research is funded through the Spanish MEC 'Ramón y Cajal' programme. AED and OG appreciate the help and support of THEMIS staff. The research by JF has been funded under MEC project CGL2005-05500-C02. The authors would like to thank T. Jahr and E. Sturkell for their comments and suggestions that greatly improved the quality of this work. Discussions with A.G. Camacho and I. Galindo have been very useful and appreciated by the authors.

References

Ablay, G. and Hurlimann, M. (2000), *Evolution of the north flank of Tenerife by recurrent giant landslides*, J. Volcanol. Geotherm. Res. *103*, 135–139.

Albert-Beltrán, J.F. Araña, V., Diez, J.L., and Valentín, A. (1990), *Physical-chemical conditions of the Teide volcanic system (Tenerife, Canary Islands)*, J. Volcanol. Geotherm. Res. *43*, 321–332.

Ancochea, E., Fúster, J.M., Ibarrola, E., Cendrero, A., Coello, J., Hernán, F., Cantagrel, J.M., and Jamond, A. (1990), *Volcanic evolution of the island of Tenerife (Canary Islands) in the light of new K-Ar data*, J. Volcanol. Geotherm. Res. *44*, 231–249.

Ancochea, E., Huertas, M.J., Cantagrel, J.M., Coello, J., Fúster, J.M., Arnaud, N., and Ibarrola, E. (1999), *Evolution of the Cañadas edifice and its implications for the origin of the Cañadas Caldera (Tenerife, Canary Islands)*, J. Volcanol. Geotherm. Res. *88*, 177–199.

Araña, V. (1985), *Evolución y mezcla de magmas en Las Cañadas del Teide. En Mecanismos eruptivos y estructuras profundas de volcanes italianos y españoles*. Reuní. Cient. CSIC-CNR 38–42.

Araña, V., Aparicio, A., García Cacho, L., and García García R. (1989), *Mezcla de magmas en la región central del Teide. En Los volcanes y la caldera del parque nacional del Teide (Tenerife, Islas Canarias)*, eds. V. Araña y J. Coello, 269–299.

Araña, V., Camacho, A.G., García, A., Montesinos, F.G., Blanco, I., Vieira, R., and Felpeto, A. (2000), *Internal structure of Tenerife (Canary Islandas) based on gravity, aeromagnetic and volcanological data*, J. Volcanol. Geotherm. Res, *103*, 43–64.

Bosshard, E. and Macfarlane, D.J. (1970), *Crustal structure of the western Canary Islands from seismic refraction and gravity data*, J. Geophys. Res. *75*, 4901–4917.

Bravo, T., Coello, J., and Bravo, J. (1976), *II Asamblea Nacional de Geodesia y Geofísica, Madrid*, Vol. Com. 2235–2244.

Cabrera, M.P., and Hernández-Pacheco, A. (1987), *Las erupciones históricas de Tenerife (Canarias) en sus aspectos vulcanológico, petrológico y geoquímica*, Rev. Mat. Proc. Geol. V, 143–182.

Carracedo, J.C. and Soler, V. (1983), *V Asamblea Nacional de Geodesia y Geofísica, Madrid*, Vol. Com. 2351–2363.

Carracedo, J.C. (1994), *The Canary Islands: An example of structural control on the growth of large oceanic-island volcanoes*, J. Volcanol. Geotherm. Res. *60*, 225–241.

Carracedo, J.C., Paterne, M., Guillou, H., Perez Torrado, F.J., Paris, R., Rodriguez Badiola, E., and Hansen, A. (2003a), *Dataciones radiometricas (C14 y K/Ar) del. Teide y el rift noroeste, Tenerife, Islas Canarias*, Estudios Geol. *59*, 15–29.

Carracedo, J.C., Guillou, H., Paterne, M., Pérez Torrado, F.J., Paris, R., and Badiola, E.R. (2003b), *Carbon-14 ages of the past 20 ka of eruptive activity of Teide volcano, Canary Islands*, EGS - AGU - EUG Joint Assembly, Abstracts from the meeting held in Nice, France, 6–11 April 2003, abstract #2627.

CARRASCO, D., FERNÁNDEZ, J., ROMERO, R., ARAÑA, V., MARTÍNEZ, A., MORENO, V., APARICIO, A., and PAGANINI, M., (2000), *First results from operational volcano monitoring in the Canary Islands*, ERS-ENVISAT Symposium, Gothenburg, Sweden, 6 pp.

DIEZ, J.L. and ALBERT, J.F. (1989), *Modelo termodinámico de la cámara magmática del Teide. En Los volcanes y la caldera del parque nacional del Teide (Tenerife, Islas Canarias)*, eds. V. Araña y J. Coello, 335–343.

DVORAK, J.J. and DZURISIN, D. (1997), *Volcano geodesy; the search for magma reservoirs and the formation of eruptive event*, Rev. Geophys. *35*, 343–384.

DZURISIN D. (2007) *Volcano deformation: New geodetic monitoring techniques*, (Springer-Verlag, Berlin Heilderberg New York), pp. 51–68 and pp. 89–95, ISBN 3-540-42642-6.

FARRUJIA, I., DELGADO, P., and BETHENCOURT, J. (1994), *Congr. Análisis y Evolución de la Contaminación de las Aguas Subterráneas*, Alcalá de Henares, Tomo II 397–416.

FEIGL, K.L. and DUPRÉ, E. (1999), *RNGCHN: A Program to calculate displacement components from dislocations in an elastic half-space with applications for modeling geodetic measurements of crustal deformation*, Comp. Geosci. *25*, 695–704.

FERNÁNDEZ, J., and RUNDLE, J.B. (1994), *Gravity changes and deformation due to a magmatic intrusion in a two-layered crustal model*, J. Geophys. Res. *99*, 2737–2746.

FERNÁNDEZ, J., RUNDLE, J.B., GRANELL, Rd.R., and YU T.T. (1997), *Programs to compute deformation due to a magma intrusion in elastic- gravitational layered Earth models*, Comput. Geosci. *23*, 231–249.

FERNÁNDEZ, J., CARRASCO, J.M., RUNDLE, J.B. and ARAÑA, V. (1999), *Geodetic methods for detecting volcanic unrest. A theoretical approach*, Bull. Volcanol. *60*, 534–544.

FERNÁNDEZ, J., ROMERO, R., CARRASCO, D., LUZÓN, F., and ARAÑA, V. (2002), *InSAR volcano and seismic monitoring in Spain. Results for the period 1992–2000 and possible interpretations*. Optics and Lasers in Engin. *37*, 285–297.

FERNÁNDEZ, J., YU, T.-T., RODRÍGUEZ-VELASCO, G., GONZÁLEZ-MATESANZ, J., ROMERO, R., RODRÍGUEZ, G., QUIRÓS, R., DALDA, A., APARICIO, A., and BLANCO, M.J. (2003), *New geodetic monitoring system in the volcanic Island of Tenerife, Canaries, Spain. Combination of InSAR and GPS techniques*, J. Volcanol. Geotherm. Res. *124*/3–4, 241–253.

FERNÁNDEZ, J., GONZÁLEZ-MATESANZ, F.J., PRIETO, J.F., STALLER, A., RODRÍGUEZ-VELASCO, G., ALONSO-MEDINA, A., and CHARCO, M. (2004), *GPS monitoring in the N-W part of the volcanic Island of Tenerife, Canaries, Spain. Strategy and results*, Pure Appl. Geophys., *161*, (7), 1359–1377, doi: 10.1007/s00024-004-2509-2.

FERNÁNDEZ, J., ROMERO, R., CARRASCO, D., TIAMPO, K.F., RODRÍGUEZ-VELASCO, G., APARICIO, A., ARAÑA, V., and GONZÁLEZ-MATESANZ, F.J. (2005), *Detection of displacements in Tenerife Island, Canaries, using radar interferometry*, Geophys. J. Internat., *160*, 33–45. doi: 10.1111/j.1365-246X.2005.02487.x.

FERNÁNDEZ, J., PRIETO, J., GONZÁLEZ, P., and T.R. GROUP (2006a), *Geodetic observation in Tenerife Island for volcano monitoring 2000–2006*, Geophys. Res. Abstr. *8*, 05757, SRef-ID:1067-7962/gra/EGU06-A-05757.

FERNÁNDEZ, J., CHARCO, M., RUNDLE, J.B., and TIAMPO, K. F. (2006b), *A revision of the FORTRAN codes GRAVV to compute deformation produced by a point magma intrusion in elastic-gravitational layered Earth models*, Comp. Geosci., *32*/2, 275–281. doi:10.1016/j.cageo.2005.06.015.

FISKE, R. S. and SHEPHERD, J. B. (1990), *Twelve years of ground-tilt measurements on the Souriere of St. Vicent, 1977–1989*, Bull. Volcanol. *52*, 227–241.

GONZÁLEZ DE VALLEJO L., CAPOTE, R., CABRERA, L., INSUA, J.M., and ACOSTA, J. (2003), *Paleoliquefaction evidence in Tenerife (Canary Islands) and posible seismotectonic sources*, Mar. Geophys. Res. *24*, 149–160.

GONZÁLEZ DE VALLEJO L., GARCÍA-MAYORDOMO, J., and INSUA J.M. (2006), *Probabilistic seismic-hazard assessment of the Canary Islands*, Bull. Seism. Soc. Am. *84*, 2040–2049.

GUILLOU, H., CARRACEDO, J.C., PARIS, R., and PÉREZ TORRADO, F.J. (2004), *Implications for the early shield-stage evolution of Tenerife from K/Ar ages and magnetic stratigraphy*, Earth Planet. Sci. Lett. *222*, 599–614.

HERNÁNDEZ, P.A., PÉREZ, N.M., SALAZAR, J.M., NAKAI, S., NOTSU, K., and WAKITA, H. (1998), *Diffuse emission of carbon dioxide, methane, and helium 3 from Teide volcano, Tenerife, Canary Islands*, Geophys. Res. Lett. *25*, 3311–3314.

IRWAN, M., KIMATA, F., FUJII, N., NAKAO, S., WATANABE, H., SAKAI, S., UKAWA, M., FUJITA, E., and KAWAI, K. (2003), *Rapid ground deformation of the Miyakejima volcano on 26–27 June, 2000 detected by kinematic GPS analysis*, Earth Planets Space *55*, 13–16.

MARINONI, L.B. and GUDMUNDSSON, A. (2000), *Dykes, faults and paleostresses in the Teno and Anaga massifs of Tenerife (Canary Islands)*, J. Volcanol. Geotherm. Res. *103*, 83–103.

MARTÍ, J., MITJAVILA, J., and ARAÑA, V. (1994), *Stratigraphy, structure and geomorphology of the Las Cañadas Caldera (Tenerife, Canary Islands)*, Geol. Mag. *131*, 715–727.

MEYERS, S.D., KELLY, B.G., and O'BRIEN, J.J. (1993), *An introduction to wavelet analysis in oceanography and meteorology: With application to the dispersion of Yanai waves*, Mon. Wea. Rev. *121*, 2858–2866.

MEZCUA, J., GALÁN, J., RUEDA, J. J., MARTÍNEZ, J., and BUFORN, E. (1990), *Sismotectónica de las Islas Canarias. Estudio del terremoto del 9 de Mayo de 1989 y su serie de réplicas*, Publ. Técnica 23, IGN, Madrid.

NEWHALL, C. and DZURISIN, D. (1988), *Historical unrest of large calderas of the world*, US. Geol. Surv. Bull. *1855*, 1108 pp.

OKADA, Y. (1985), *Surface deformation due to shear and tensile faults in a half-space*, Bull. Seismol. Soc. Am. *75*,1135–1154.

PÉREZ, N.M., NAKAI, S., WAKITA, H., HERNÁNDEZ, P.A., and SALAZAR, J.M. (1996), *Helium-3 emission in and around Teide volcano, Tenerife, Canary Islands, Spain*, Geophys. Res. Lett. *23*, 3531–3534.

PÉREZ, N.M., MELÍAN G., GALINDO, I., PADRÓN, E., HERNÁNDEZ, P.A., NOLASCO, D., SALAZAR, P., PÉREZ, V., COELLO, C., MARRERO, R., GONZÁLEZ, Y., GONZÁLEZ, P. and BARRANCOS, J. (2005), *Premonitory geochemical and geophysical signatures of unrest at Tenerife, Canary Islands*, Geophys. Res. Abstracts *7*, 9993.

ROMERO, R., FERNÁNDEZ, J., CARRASCO, D., LUZÓN, F., MARTÍNEZ, A., RODRÍGUEZ-VELASCO, G., MORENO, V., ARAÑA, V., and APARICIO, A. (2002), *Synthetic Aperture Radar Interferometry (InSAR): Application to ground deformations studies for volcano and seismic monitoring*, Física de la Tierra *14*, 55–84.

ROMERO-RUIZ, C., GARCÍA-CACHO, L., ARAÑA, V., YANES LUQUE, A., and FELPETO, A. (2000), *Submarine volcanism surrounding Tenerife, Canary Islands: Implications for tectonic controls, and oceanic shield forming processes*, J. Volcano. Geotherm. Res. *103*, 105–119.

RUNDLE, J.B. (1982), *Deformation, gravity, and potential changes due to volcanic loading of the crust*, J Geophys. Res. *87*(10), 729–744.

SALEH, B., BLUM, P.A., and DELORME H. (1991), *New silica compact tiltmeter for deformations measurements*, J. Surv. Engin. ASCE, *117* (1), 27–35.

SEVILLA, M.J. and MARTÍN M.D. (1986), *Geodetic network design for crustal deformation studies in the Caldera of Teide area*, Tectonophysics *130*, 235–248.

SEVILLA, M.J., VALBUENA, J.L., RODRÍGUEZ-DÍAZ, G., and VARA, M.D. (1996), *Trabajos altimétricos en la Caldera del Teide*, Física de la Tierra *8*, 117–130.

TORRENCE, C. and COMPO, G.P. (1998), *A Practical Guide to Wavelet Análisis*, Bull. Amer. Meteor. Soc. *79*, 61–78.

VALENTÍN, A., ALBERT-BELTRÁN, J., and DIEZ, J. (1990), *Geochemical and geothermal constraints on magma bodies associated with historic activity, Tenerife (Canary Islands)*, J. Volcanol. Geotherm. Res. *44*, 251–264.

YAMAOKA, K., KAWAMURA, M., KIMATA, F., FUJII, N., and KUDO, T. (2005), *Dike intrusion associated with the 2000 eruption of Miyakejima Volcano*, Japan. Bull. Volcanol *67*, 231–242.

YU, T. T., FERNÁNDEZ, J., TSENG, C.L., SEVILLA, M.J., and ARAÑA, V. (2000), *Sensitivity test of the geodetic network in Las Cañadas Caldera, Tenerife, for volcano monitoring*, J. Volcanol. Geotherm. Res. *103*, 393–407.

(Received July 24, 2007, revised March 28, 2008, accepted April 3, 2008)
Published Online First: July 14, 2008

Pure appl. geophys. 165 (2008) 1071–1093
0033–4553/08/061071–23
DOI 10.1007/s00024-008-0347-3

Pure and Applied Geophysics

Multi-Channel Satellite Image Analysis Using a Variational Approach

L. Alvarez, C. A. Castaño, M. García, K. Krissian, L. Mazorra, A. Salgado and J. Sánchez

Abstract—Currently, meteorological satellites provide multichannel image sequences including visible, temperature and water vapor channels. Based on a variational approach, we propose mathematical models to address some of the usual challenges in satellite image analysis such as: *(i)* the estimation and smoothing of the cloud structures by decoupling them into different layers depending on their altitudes, *(ii)* the estimation of the cloud structure motion by combining information from all the channels, and *(iii)* the 3D visualization of both the cloud structure and the estimated displacements. We include information of all the channels in a single variational motion estimation model. The associated Euler-Lagrange equations yield to a nonlinear system of partial differential equations that we solve numerically using finite-difference schemes. We illustrate the performance of the proposed models with numerical experiments on two multichannel satellite sequences of the North Atlantic, one of them from the Hurricane Vince. Based on a realistic synthetic ground truth motion, we show that our multichannel approach overcomes the single channel estimation for both the average Euclidean and angular errors.

Key words: Variational methods, Partial Differential Equations, optical flow, spatio-regularization, optimization, satellites images.

1. Introduction

Estimation of the cloud motion from satellite images has several applications in meteorology and climate (Hasler, 1990). In particular, it is an important source of information for numerical weather prediction (NWP) (Baker, 1991). It is also useful in understanding the structure and dynamics of hurricanes and severe thunderstorms.

Different classes of techniques have been used to estimate the cloud motion, among them are techniques using the local cross correlation (Leese *et al*., 1971; Phillips *et al*., 1972; Schmetz *et al*., 1993) and cross-correlation combined with relaxation labeling (Wu, 1995; Evans, 2006), motion analysis from stereoscopic images (Young and Chellappa, 1990; Kambhamettu *et al*., 1995), neural networks (Côté and Tatnall 1995), block-matching techniques (Brad and Letia, 2002), local fitting (Zhou *et al*., 2001), scale-space classification by matching contour points of high curvatures (Mukherjee and Acton,

Dpto. de Informática y Sistemas, Universidad de Las Palmas de Gran Canaria, Campus Universitario de Tafira s/n, 35017 Las Palmas de Gran Canaria, Spain. E-mail: lalvarez@dis.ulpgc.es; http://serdis.dis.ulpgc.es/lalvarez/ami/index.html

2002), and variational techniques using Partial Differential Equations (PDE) also referred to as Optical Flow techniques in the field of computer vision (CORPETTI *et al*., 2002). The most widely used techniques are based on the local cross correlation. They have the advantage of being robust to global intensity changes, but they are also computationally expensive and they do not integrate a global regularization constraint. In this context, regularization means including in the estimation model a constraint which ensures spatial coherence of the results, that is, the motion field should be a smooth function. In correlation-based techniques, the displacement motion is usually calculated on a discrete regular lattice of the image domain. Also, local cross-correlation techniques are well suited for rigid-motion but can fail in case of fast non-rigid displacements. Conversely, methods based on variational optical flow impose a global smoothing constraint on the estimated displacement and are calculated on the whole image domain, leading to a dense estimation of the flow.

The tracking of cloud motion is usually computed from the infrared (IR: 10.5–12.5 μm) channel (LEESE *et al*., 1971; SCHMETZ *et al*., 1993). Low-level cloud motion has also been tracked from the visible channel (OTTENBACHER *et al*., 1997; ZHOU *et al*., 2001). Recently, multispectral motion analysis has been investigated, first by visual superimposition of the cloud motion estimated on each channel individually (VELDEN *et al*., 1998) and more recently using a multichannel cross-correlation technique with the visible and IR channels (EVANS, 2006). Recently, HÉAS *et al*. (2006) have also proposed a dense multi-layer estimation of the cloud motion.

In this work, we introduce several models based on a variational approach to analyze multichannel satellite images. First, we propose an energy minimization technique to address the problem of cloud structure motion estimation, we combine the information of several channels in order to improve the accuracy of the cloud motion estimation. Next, we address the problem of cloud structure layer classification and channel smoothing. Classifying the cloud structures in different layers following the different altitudes of the clouds is an important issue. The meteorological satellites provide an initial layer classification based on a combination of the channels. This initial layer classification is noisy because of the semi-transparency of some clouds, covering part of the pixel area (of 3×3 km in our images) and because of the noise introduced during the image acquisition. Therefore, it requires some smoothing. We present a variational technique based on a median-type filter to smooth the boundary of the initial cloud classification layers. We also propose, using this layer classification, a technique to smooth the different channels which works independently in each layer. The advantage of this technique is to preserve the discontinuities of the channel signal across the boundaries of the different layers.

The organization of the paper is as follows: in section 2, we describe the images obtained from Meteosat Second Generation (MSG) satellite and we also describe the initial optical flow technique (ALVAREZ et al., 2000) that will be the base of our multichannel sequential motion tracking algorithm. In section 3, we explain how we extend the variational optical flow method to deal with multichannel sequential data.

In section 4, we present a technique to smooth the boundary of the cloud classification layers and the channel signals. In section 5, we describe a procedure for three-dimensional visualization of the data, including the different classes of clouds and the motion vectors. Section 6 shows numerical experiments on two satellite image sequences and finally, we conclude in section 7.

2. Background

In this section, we introduce the two satellite image sequences that we used in our experiments and we also describe the variational optical flow technique that will be later extended to deal with multichannel sequential images.

2.1. Satellite Images

Meteosat Second Generation satellites replaced in 2002 the former Meteosat, providing a significantly increased amount of information as compared to the previous version in order to continuously observe the whole Earth. In this sense, MSG generates images every 15 min with a 10-bit quantization, a spatial sampling distance of 3 km at subsatellite point in 11 channels, from the visible to the infrared channel, and 1 km in the high resolution visible channel (SCHMETZ *et al*., 2002). All these channels provide information that is used for different applications, summarized in Table 1. Among the most important applications, numerical weather prediction combines the information from different channels, mainly from the VIS 0.8, WV 6.2, WV 7.3 and IR 10.8 channels (SCHMETZ *et al*., 2002), to compute the displacement of the clouds between two time instants, that constitute the most important source of information for this application.

Table 1

Characteristics and main applications of the different MSG channels

Name of the Channel	Central Wavelength	Main Application
VIS 0.6	0.63 μm	Cloud detection and tracking, surface identification
VIS 0.8	0.81 μm	Cloud detection and tracking, surface identification
NIR 1.6	1.64 μm	Discrimination snow/ice cloud/water cloud
IR 3.9	3.92 μm	Detection of low cloud/fog at night
WV 6.2	6.25 μm	Water vapor structures at medium-high level
WV 7.3	7.35 μm	Water vapor structures at low-medium level
IR 8.7	8.70 μm	Ice/water distinction
IR 9.7	9.66 μm	Ozone detection in lower stratosphere
IR 10.8	10.80 μm	Estimation of temperature of clouds and surface
IR 12	12.00 μm	Estimation of temperature of clouds and surface
IR 13.4	13.40 μm	Cloud height estimation
HRV	0.75 μm	High spatial resolution

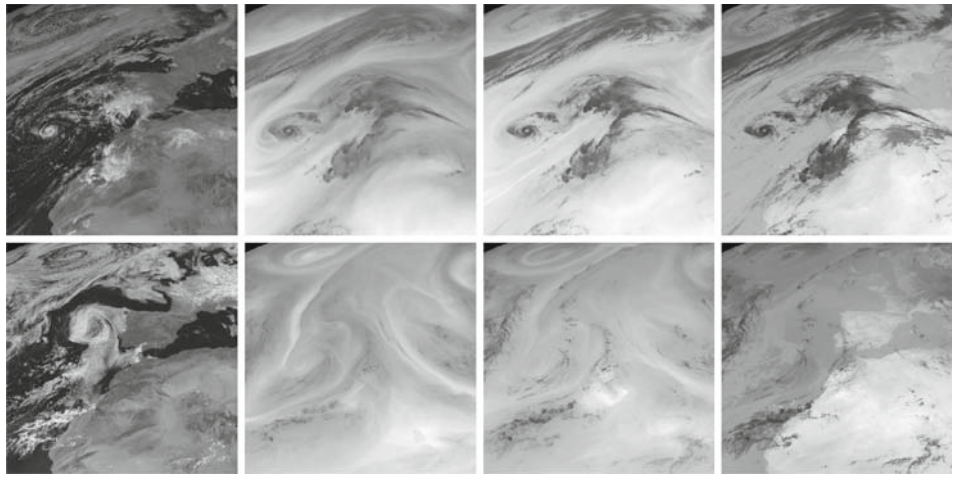


Figure 1
Images from different channels of the MSG satellite. From left to right, respectively, the VIS 0.8, WV 6.2, WV 7.3 and IR 10.8. Top row: Images from the Vince hurricane. Bottom row: Images from the sequence on June 5, 2004. Both sequences are from the North Atlantic area.

The VIS 0.8 channel provides information on the visible zone of the spectrum, that allows identification of cloud structures in the atmosphere, cloud tracking and land and vegetation monitoring. Water Vapor channels, WV 6.2 and WV 7.3, allow us to observe water vapor and winds, and also support height allocation of semi-transparent clouds (SCHMETZ *et al*., 1993). Finally, IR 10.8 is essential for measuring temperatures at sea and land surface and the top of clouds, and detection of cirrus clouds (INOUE, 1987). Another important application, used in the visualization section 4, is cloud structure classification which consists in an estimation of the cloud structure altitude using this multichannel information, yielding to a segmentation of the pixels into different types of clouds (SZANTAI and DÉSALMAND, 2005; AMEUR *et al*., 2004), as shown in Figure 2.

2.2. Optical Flow Techniques in Computer Vision

Numerous methods have been proposed in the computer vision community to address the problem of motion estimation from a set of images. The projection of the 3D object motion in the scene yields a 2D flow field in the image domain. Most of the methods deal with the problem of estimating the 2D vector field between images based on the image intensities. This problem is generally referred to as "optical flow estimation." The optical flow is the apparent displacement of pixels in a sequence of images.

During the last two decades many techniques for computing the optical flow have appeared. These methods can be classified into three different categories: correlation–based, gradient–based and phase–based techniques (BEAUCHEMIN and BARRON, 1995;

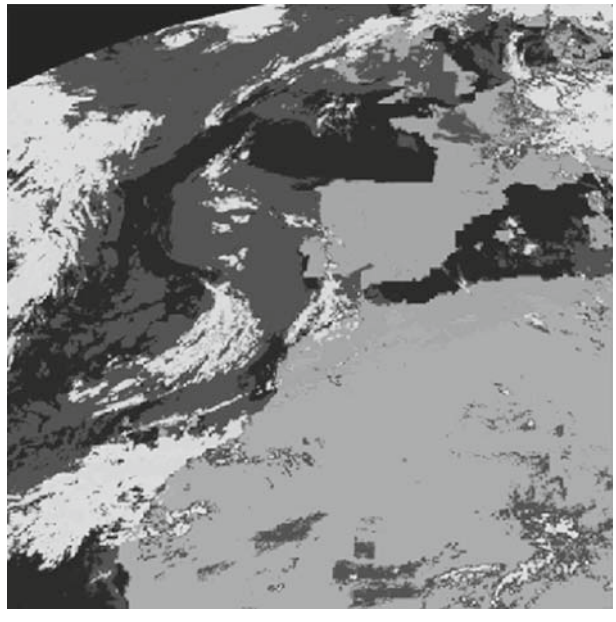


Figure 2
In this image, we illustrate, using different greyscale values, the original cloud structure layer classification estimated from the meteorological satellite channels.

Mitiche and Bouthemy, 1996). Different works have also evaluated the performance of the most popular algorithms (Barron *et al*., 1994; Jähne and Haussecker, 2000; Galvin *et al*., 1998). The gradient–based techniques are amongst the most accurate and robust strategies to calculate the 2D flow field. They rely on the so–called *optical flow constraint* which relates to the brightness gradient with the vector field, $\mathbf{h}(\mathbf{x}) = (u(\mathbf{x}),v(\mathbf{x}))$.

The determination of optical flow is a classic ill–posed problem in computer vision and it requires additional regularizing assumptions. The regularization by Horn and Schunck (1980) reflects the assumption that the optical flow field varies smoothly in space. However, since many natural image sequences are better described in terms of piecewise smooth flow fields, much research has been done to modify the Horn and Schunck approach to permit discontinuous flow fields (Nagel and Enkelmann, 1986; Proesmans *et al*., 1994; Aubert *et al*., 1999; Black and Anandan, 1996; Deriche *et al*., 1995; Weickert and Schnörr, 2001; Papenberg *et al*., 2006).

2.3. Variational Formulation

The 2D flow computation is carried out using a PDE-based optical flow technique described in Alvarez *et al*. (2000). It consists in minimizing an energy defined as a weighted sum of two terms: A data term and a regularization term. The data term assumes

that the images have similar intensities at the corresponding points and the regularization term assumes smoothness of the fluid flow.

The regularization term uses the approach proposed by NAGEL and ENKELMANN (1986), with the following improvements: (i) The formulation avoids inconsistencies caused by centering the brightness term and the smoothness term in different images, (ii) it uses a coarse to fine linear scale-space strategy to avoid convergence to physically irrelevant local minima, and (iii) it creates an energy functional that is invariant under linear brightness changes.

The energy to minimize is written as:

$$E(\mathbf{h}) = \int_{\mathbb{R}^2} (I_1(\mathbf{x}) - I_2(\mathbf{x}+\mathbf{h}))^2 d\mathbf{x} + C \int_{\mathbb{R}^2} \mathbf{tr}(\nabla \mathbf{h}^{\mathbf{t}} \mathbf{D} \nabla \mathbf{h})\, \mathbf{dx}, \tag{1}$$

where $\mathbf{x}$ is a point in $\mathbb{R}^2, \mathbf{h} = \mathbf{h}(\mathbf{x}) = (\mathrm{u}(\mathrm{x}), \mathrm{v}(\mathrm{x}))^{\mathrm{t}}$ is the displacement field that we are looking for, $(.)^t$ is the transpose operator, I_1 and I_2 are the two input images, tr(.) is the trace operator, C is a constant that weights the smoothing term, ∇ is the gradient operator, and D is a regularized projection matrix in the direction orthogonal to ∇I_1.

The matrix D is expressed as:

$$D(\nabla I_1) = \frac{1}{\|\nabla I_1\|^2 + 2\lambda^2} \left(\xi \xi^t + \lambda^2 Id\right), \tag{2}$$

where $\xi = (\partial I_1/\partial y, -\partial I_1/\partial x)^t$ is a vector orthogonal to $\nabla I_1, \|.\|$ denotes the norm operator, Id is the identity matrix, and λ is a coefficient that determines the isotropic behavior of the smoothing and inhibits blurring across the boundaries for gradients of high magnitude: $\|\nabla I_1\| \gg \lambda$.

C and λ are computed by means of two parameters α and $s \in (0, 1)$ where

$$C = \frac{\alpha}{\max\left(|(\nabla G_\sigma * I_1)(\overline{x})|^2\right)}, \tag{3}$$

$$s = \int_0^\lambda H_{|\nabla G_\sigma * I_1|}(z) dz \tag{4}$$

where G_σ * I_1 represents the convolution of I_1 with a Gaussian of standard deviation $\sigma, H_{|\nabla G_\sigma * I_1|}(z)$ represents the normalized histogram of $|\nabla G_\sigma * I_1|$. s is called the *isotropy fraction*. When $s \to 0$, the diffusion operator becomes anisotropic whereas $s \to 1$, it leads to isotropic diffusion. This normalization of C and λ allows the energy to be invariant under grey level transformation of the form $(I_1, I_2) \to (kI_1, kI_2)$. In the tests, the input parameters are α and s. C and λ are computed as in equations (3) and (4). Thus, these parameters are automatically adjusted to the dynamic image range. See ALVAREZ *et al*. (2000) for more details.

The associated Euler-Lagrange equations are given by the following PDE system:

$$C\,\mathrm{div}(D\nabla u) + (I_1(\mathbf{x}) - I_2(\mathbf{x}+\mathbf{h}))\frac{\partial I_2}{\partial x}(\mathbf{x}+\mathbf{h}) = 0 \tag{5}$$

$$C\,\mathrm{div}(D\nabla v) + (I_1(\mathbf{x}) - I_2(\mathbf{x}+\mathbf{h}))\frac{\partial I_2}{\partial y}(\mathbf{x}+\mathbf{h}) = 0 \tag{6}$$

The system is numerically solved using an iterative Gauss-Seidel algorithm as detailed in ALVAREZ *et al*. (2000).

3. Multi-channel Flow Computation

The variational methods proposed in the literature usually use information from a single channel. Normally, these methods are targeted at solving the optic flow problem using greyscale visual images (one channel). Although methods oriented to color (multi-channel) images have been investigated, they are not very common. The satellites have many sensors that capture images using different regions of the wave spectrum. This information is very useful for the estimation of the cloud motion. The clouds produce changes in the water vapor concentration, in the air pressure and in the thermal radiation from the earth. Consequently, using the data given by these channels, we should be able to compute a more robust and more accurate solution of the cloud motion. We have extended the optical flow method described in section 2.3 to include information from several channels captured by the satellite sensors. In this section, we explain in detail our variational (energy) model and the corresponding numerical scheme.

3.1. Variational Formulation

Energy model. The energy model proposed for motion estimation using multichannel data is, as in the case of a single channel, based on the addition of two terms: the data term and the smoothness term. Our input data will be pairs of images of different channels. Our energy has to combine the information from different channels. We have included in the energy a set of weights that specifies the relevance of each channel. The data term takes into account information from all the channels. Thus, it constitutes in a combination of differences between two images weighted by positive real numbers ρ_c, where $c \in [1, N_c]$ is the channel associated with this number and N_c is the number of channels. In the single channel method, the Nagel-Enkelmann operator uses the image gradient to decide the direction and the amount of diffusion. In the multi-channel method, we want to keep this idea by combining the data from different channels.

The energy to minimize is written as:

$$E(\mathbf{h}) = \int_{\mathbb{R}^2} \sum_{c=1}^{N_c} \rho_c (I_1^c(\mathbf{x}) - I_2^c(\mathbf{x}+\mathbf{h})^2 d\mathbf{x} + C \int_{\mathbb{R}^2} \mathbf{tr}(\nabla \mathbf{h}^{\mathbf{t}} \mathbf{D} \nabla \mathbf{h}) d\mathbf{x}, \tag{7}$$

where I_1^c and I_2^c are the first and second images in the channel c. We follow the same notation as section 2.3.

In the single channel method the matrix D is expressed as:

$$D(\nabla I) = \frac{1}{\|\nabla I\|^2 + 2\lambda^2}\left(\xi\xi^t + \lambda^2 Id\right), \tag{8}$$

where $\xi = (\partial\nabla I/\partial y, -\partial\nabla I/\partial x)^t$ is a vector orthogonal to ∇I. In order to define matrix D for the multi-channel method, we need to define a single vector $\bar{g}$ which plays the role of ∇I. To define $\bar{g}$, we propose two strategies:

– *Maximum gradient*. At each pixel location, $\bar{g}$ is computed as the gradient of greatest magnitude among the gradients of all the channels:

$$\operatorname{argmax}\left\{\|\overrightarrow{v}\|, \overrightarrow{v} \in \{\nabla I^c, c \in [1, N_c]\}\right\}$$

– *Average gradient*. $\bar{g}$ is computed as a dominant direction in the set of the gradient vectors for all channels. The usual way to estimate the dominant orientation $\bar{g}$ is using the so-called structure tensor. This structure tensor is defined as the matrix

$$\sum_{c=1}^{N_c} \rho_c\left(\nabla I_i^c\right) \cdot \left(\nabla I_i^c\right)^t.$$

If we denote $\bar{e}_{\max}$ the normalized eigenvector associated with the maximum eigenvalue $\lambda_{\max}$ of the above matrix, then we can define $\bar{g}$ as:

$$\bar{g} = \sqrt{\frac{\lambda_{\max}}{\sum_{c=1}^{N_c} \rho_c}}\bar{e}_{\max}.$$

In fact, we can show that $\bar{g}$ is the minimum, under the constraint $\|\bar{g}\| = \sqrt{\lambda_{\max}/\sum_{c=1}^{N_c} \rho_c}$, of the following energy :

$$E(\bar{g}) = -\sum_{c=1}^{N_c} \rho_c\left(\bar{g}^t \cdot \nabla I_i^c\right)^2.$$

3.2. Numerical Scheme

The Euler-Lagrange equations associated with the energy (7) are:

$$C\,\mathrm{div}(D\nabla u) + \sum_{c=1}^{N_c} \rho_c(I_1^c(\mathbf{x}) - I_2^c(\mathbf{x}+\mathbf{h}))\frac{\partial I_2^c}{\partial x}(\mathbf{x}+\mathbf{h}) = 0 \tag{9}$$

$$C\,\mathrm{div}(D\nabla v)+\sum_{c=1}^{N_c}\rho_c(I_1^c(\mathbf{x})-I_2^c(\mathbf{x}+\mathbf{h})\frac{\partial I_2^c}{\partial y}(\mathbf{x}+\mathbf{h})=0 \tag{10}$$

To discretize the above system of partial differential equations, we use an implicit finite-difference scheme because it is more stable and converges faster than the usual explicit schemes.

The matrix D, introduced above, is written in each pixel i as:

$$D_i=\begin{pmatrix}a_i & b_i\\ b_i & c_i\end{pmatrix}$$

We can discretize the differential operator in each pixel i, and we obtain:

$$\mathrm{div}(D_i\nabla u)=\begin{pmatrix}a_i\partial_x u+b_i\partial_y u\\ b_i\partial_x u+c_i\partial_y u\end{pmatrix}=\partial_x(a_i\partial_x u)+\partial_x\big(b_i\partial_y u\big)+\partial_y(b_i\partial_x u)+\partial_y\big(c_i\partial_y u\big)$$

We define N_i^* as the set of 3×3 neighbors around the pixel i excluding the pixel i itself. Using a standard difference scheme we can write:

$$\mathrm{div}(D_i\nabla u_i)=\sum_{n\in N_i^*}(d_n u_n)+d_i u_i \tag{11}$$

for suitable coefficients d_n. The same expression with the same coefficients can be obtained for $\mathrm{div}(D_i\nabla\ v)$.

The components of the vector displacement (u_i,v_i) are obtained asymptotically by iterations of a Gauss-Seidel type scheme, where k denotes the iteration number. To this end, the terms $I(\mathbf{x}+\mathbf{h}^{k+1})$ are linearized by Taylor expansion:

$$I_1^c(\mathbf{x})-I_2^c(\mathbf{x}+\mathbf{h}^{k+1})\simeq I_1^c(\mathbf{x})-I_2^c(\mathbf{x}+\mathbf{h}^k)-\frac{\partial I_2^c}{\partial x}(\mathbf{x}+\mathbf{h}^k)(u^{k+1}-u^k)-\frac{\partial I_2^c}{\partial y}(\mathbf{x}+\mathbf{h}^k)(v^{k+1}-v^k).$$

Denoting $I_{2,i,x}^{c,k}=\partial I_2^c/\partial x(\mathbf{x}_i+\mathbf{h}^k)$ and $I_{2,i,y}^{c,k}=\partial I_2^c/\partial y(\mathbf{x}_i+\mathbf{h}^k)$, we finally obtain:

$$u_i^{k+1}=\frac{u_i^k+dt\left(C\sum\limits_{n\in N_i^*}\big(d_n u_n^k\big)+\sum\limits_{c=1}^{N_c}\rho_c\Big(I_1^c(\mathbf{x}_i)-I_2^c(\mathbf{x}_i+\mathbf{h}^k)+u_i^k I_{2,i,x}^{c,k}-(v_i^{k+1}-v_i^k)I_{2,i,y}^{c,k}\Big)I_{2,i,x}^{c,k}\right)}{1+dt\left(Cd_i+\sum\limits_{c=1}^{N_c}\rho_c\Big(I_{2,i,x}^{c,k}\Big)^2\right)}$$

$$v_i^{k+1}=\frac{v_i^k+dt\left(C\sum\limits_{n\in N_i^*}\big(d_n v_n^k\big)+\sum\limits_{c=1}^{N_c}\rho_c\Big(I_1^c(\mathbf{x}_i)-I_2^c(\mathbf{x}_i+\mathbf{h}^k)+v_i^k I_{2,i,y}^{c,k}-(u_i^{k+1}-u_i^k)I_{2,i,x}^{c,k}\Big)I_{2,i,y}^{c,k}\right)}{1+dt\left(Cd_i+\sum\limits_{c=1}^{N_c}\rho_c\Big(I_{2,i,y}^{c,k}\Big)^2\right)}$$

In order to increase the convergence rate of the algorithm and to avoid being trapped in spurious local minima, we use a multi-resolution scheme, that is, we solve

successively the system at different levels of image resolution, starting from the coarsest grid.

4. Cloud Structure Smoothing

Meteorological satellites provide a cloud structure classification based on an estimation of the cloud structure altitude computed using a combination of the multi-channel satellite image values (see Fig. 2 for illustration). In practice, we can assume that the classification areas L_i ($i = 1,\ldots,N_L$) are estimated as a level set of as a classification function f_L: $\Omega \to R$, that is

$$L_i = \{\bar{x} \in \Omega : \beta_{i-1} < f_L(\bar{x}) < \beta_i\},$$

where $\beta_0 < \beta_1 < \ldots\ldots < \beta_{N_L}$. Each classification area L_i represents a cloud structure layer. Usually there are two main problems concerning the classification areas L_i. The first one is that the multi-channel satellite image values are noisy and therefore the classification function $f_L(.)$ is also noisy and the layers L_i require some kind of smoothing. The second one is that, in order to analyze the cloud structures, we need to assume a model of interaction between the different layers L_i. In this paper, we will assume the simplest case each layer L_i is at a different altitude and there is no interaction between the different layers. This assumption is usually true, but it obviously fails in the case of complex 3D atmospheric phenomena as, for instance, the hurricanes.

In order to smooth the boundary of the cloud layers L_i, we propose to use a median type filter applied to the classification function $f_L(.)$ that is, for each $\bar{x}$, we define $\mathrm{med}(f_L)(\bar{x})$ as:

$$\mathrm{med}(f_L)(\bar{x}) = \left\{ \beta : \int_{B(\bar{x})} |f_L(\bar{y}) - \beta| d\bar{y} \leq \int_{B(\bar{x})} |f_L(\bar{y}) - v| d\bar{y} \quad v \in R \right\}$$

where $B(\bar{x})$ is a neighborhood of $\bar{x}$. Thus we define the new classification layers $L_i^{'}$ as

$$L_i' = \{\bar{x} \in \Omega : \beta_{i-1} < \mathrm{med}(f_L)(\bar{x}) < \beta_i\}$$

this filter smoothes the boundary of the layers L_i and it removes the small isolated set in L_i, as depicted in Figure 3.

In order to smooth the multi-channel satellite image values $I_j^c(.)$, we use a variational technique based on the following energy minimization

$$E(u_j^c) = \int_{L_i'} \left(u_j^c(\bar{y}) - I_j^c(\bar{y}) \right)^2 d\bar{y} + \alpha \int_{L_i'} \left\| \nabla u_j^c(\bar{y}) \right\|^2 d\bar{y},$$

the parameter α represents the weight of the regularization process. Since we assume no interaction between the different layers $L_i^{'}$, we will consider a homogeneous Neumann

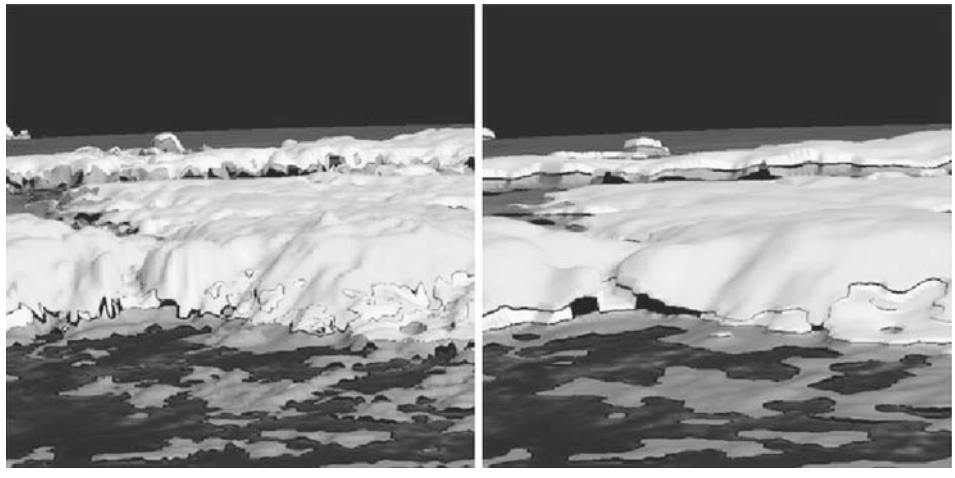


Figure 3
On the left, we present the original cloud structure layer where the altitude is computed from the channel IR 10.8. On the right, we present the results obtained with the proposed methods. We observe the smoothing of the cloud structure layer boundary as well as the smoothing of the height.

type boundary condition. The Euler-Lagrange equation associated to the above energy is given by :

$$\begin{cases} -\alpha \Delta u_j^c(\bar{x}) + u_j^c(\bar{x}) = I_j^c(\bar{x}) & \text{in} \quad L_i' \\ \frac{\partial u_j^c}{\partial \bar{n}}(\bar{x}) = 0 & \text{in} \quad \partial L_i' \end{cases},$$

The solution $u_j^c(.)$ of the above differential equation represents the smoothed version of the channel $I_j^c(\bar{x})$ in the classification area $L_i^{'}$. We observe that, since we smooth $I_j^c(.)$ in an independent way in each classification area $L_i^{'}$, the discontinuities across the boundary of the cloud layers $L_i^{'}$ of the satellite image value $I_j^c(.)$ are preserved. Figure 3 illustrates this behavior. We observe the smoothing of the cloud structure layer boundary as well as the smoothing of the height.

5. Visualization

The 3D visualization is based on OpenGL, using our software AMILab[1]. Figure 4 illustrates the different tasks and their inputs.

The height of the clouds is computed using the technique described in SZANTAI and DÉSALMAND (2005), chapter 4: "Height assignment of motion vectors." An approximation of the height of the clouds is computed from an estimation of the temperature based on

[1] http://serdis.dis.ulpgc. es/∼krissian/HomePage/Software/AMILab/

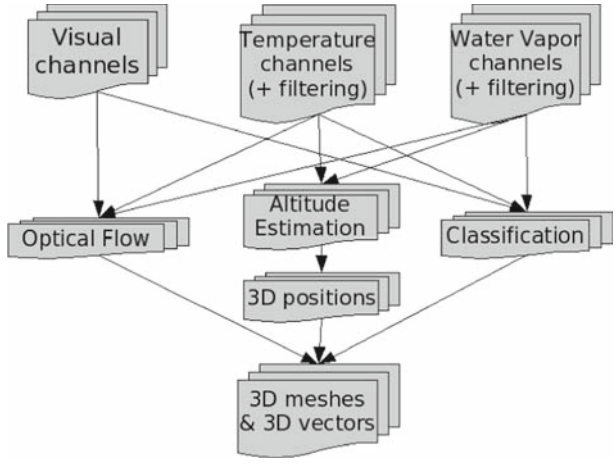


Figure 4
Processing and 3D visualization of satellite data

the IR channel 10.8. Let us denote the infrared intensity at the current pixel position as C. The *radiance* R (mW.m^{-2} . sr^{-1} . cm) is calculated as $R = R_0 + \alpha C$, where R_0 and α are included in the original MSG files. The *brightness temperature* T_b of the observed object can be approximated from the infrared channel, using the formula:

$$T_b = \frac{1}{A}\left(\frac{(C_2 v_c)}{\ln(1+\frac{C_1 v_c^3}{R})} - B\right), \tag{12}$$

where $C_1 = 1.19104.10^{-}5$ mW.m^{-2}.sr^{-1}. cm^4, $C_2 = 1.43877$ K.cm, and the central wavenumber v_c, the parameters A (dimensionless) and B (in Kelvin) are constants given by Eumetsat for each satellite and channel. The *altitude* is then deduced from the temperature as:

$$a = \frac{T - T_0}{\gamma}, \tag{13}$$

where $T_0 = 288.15$K (15°C) is the approximate temperature at sea level, and $\gamma = -6.5\ 10^{-3}$K.m^{-3} is the standard temperature change with respect to the height.

Figure 5 illustrates our 3D visualization. Each pixel of the scene is drawn at its 3D location based on its estimated height and on the parameters of the satellite. Using the cloud classification provided by Eumetsat, low, medium, high and very high clouds are displayed with different greyscale values. The estimated cloud motion is represented by 3D vectors which are displayed on a regular grid within the identified clouds. The 3D vectors are proportional to the estimated displacement and have a vertical component that represents the evolution of the clouds height. This component is based on the height estimation of two successive frames of the sequence, and on the 2D motion

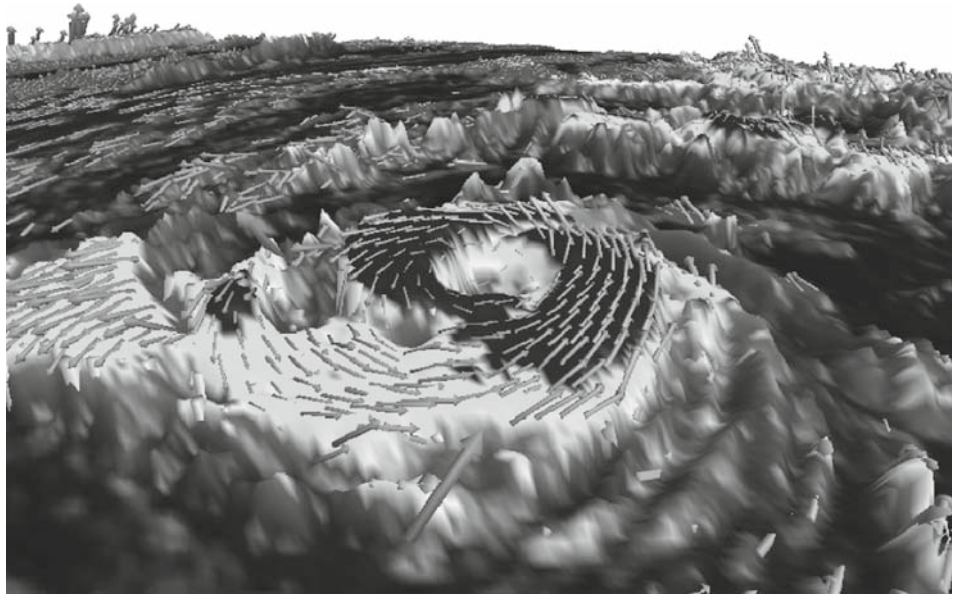


Figure 5
3D view of the hurricane Vince, including the clouds and 3D displacement vectors calculated from our multi-channel motion estimation algorithm.

field. A video that illustrates our results is also available on internet (ALVAREZ *et al.*, 2006)[2].

6. Experimental Results

In this section, we present the experiments performed using the variational multi-channel method presented in section 3 on two satellite sequences.

Since the real cloud motion is unknown for the multi-channel satellite sequences, we have designed a synthetic experiment to compare the different methods. This synthetic experiment is based on the real datasets and its generation is divided in the following steps:

1. For each channel, and for four consecutive frames in the satellite sequence, we compute a cloud motion estimation using correlation-based techniques. Thus, we will obtain twelve different motion fields (3 motion fields for each channel).
2. We build a synthetic motion model taking at each pixel location, the median (for each coordinate) of the twelve motion field values obtained for the different channels across

[2] http://serdis.dis.ulpgc. es/∼krissian/HomePage/Demos/Fluid/Video/CVPR_VIDEO_AMI.mpg

the sequence. Since the synthetic motion model is obtained from estimation on a real satellite image sequence, we expect it to be realistic. We used all channels at different times and a correlation-based technique to estimate the cloud motion, so we consider that the obtained synthetic model is not biased neither towards a particular channel, nor a particular variational model. For these reasons, we believe that this model is appropriate to compare different variational techniques.

3. A synthetic satellite sequence has been computed by warping the real satellite sequence according to the synthetic motion model. For this synthetic satellite sequence, the motion ground truth is the synthetic motion model, and a quantitative comparison between different estimation algorithms can be performed.

The multichannel satellite data are used in different ways depending on the application. The clouds are easily observable in the visible channel while the infrared channels provide other types of useful and complementary information.

The weights ρ_c, defined in (7), determine the importance of each channel. In order to compare the amount of motion contributed by each channel in the multi-channel solution, we also computed the flow in each channel separately.

In our tests, we used two satellite sequences from the North Atlantic area. The first sequence corresponds to Hurricane Vince (October 8, 2005) and the second sequence is from June 5, 2004. From these real sequences, we generate synthetic ones using the above described procedure. The new synthetic sequences satisfy the following properties:

– the pixel displacement is known,
– the motion is not biased toward a particular channel because the information of all channels is used in the same way,
– the motion is not biased toward a particular variational method because the synthetic motion model is obtained from cross-correlation techniques.

To evaluate the quality of our variational method, we used two error measures: the Average Euclidean Error (AEE, eq. 14) and the Average Angular Error (AAE, eq. 15). See BARRON *et al*. (1994) for more details.

$$AEE = \frac{1}{N}\sum_{i=1}^{N} |\mathbf{u}_i - \mathbf{u}_{ref\,i}|, \tag{14}$$

$$AAE = \frac{180}{N\pi}\sum_{i=1}^{N} \arccos\left(\frac{\mathbf{u}_i\mathbf{u}_{ref\,i} + 1}{\sqrt{\|u_i\|^2 + 1}\sqrt{\|u_{ref\,i}\|^2 + 1}}\right), \tag{15}$$

where N is the number of pixels in the image, $\mathbf{u}_i$ and $\mathbf{u}_{\mathrm{ref}\,i}$ are the flow fields in the pixel i in the computed image and the ground truth, respectively.

In the experimental results, we compare two variational methods: SF (*Simple Flow*) and MC (*Multi-Channel*). The main difference between them is that SF uses as input data

a pair of images (single channel data) whereas the input data of MC is a set of a pair of images (multi-channel data). In the comparison, we show two different versions of the MC method. They correspond to the two criteria: *Maximum gradient* and *Average gradient*, described in section 3.1. In the experiments, the SF method is applied to each channel in order to show the contribution of each channel separately, whereas the MC method is applied to the four channels. Results are presented in Tables 2 and 3, comparing both methods. We output, from top to bottom, the four estimations obtained with SF the method (one per channel) and the two estimations obtained with the MC method and from left to right, the two quantitative error measures described previously.

In Figures 6 and 9, we show the solution obtained using each channel separately, for each of the two sequences. Based on our experience, more coherent and robust results are obtained by assigning a higher weight to the visible channel. In the experiments, we use the weights 0.5 for the visible channel, 0.3 for infrared channel and 0.1 for each other channel in Vince sequence and [0.5, 0.1, 0.2, 0.2] = (VIS, VP1, VP2, IR) for North Atlantic sequence. Giving high weighting values to the vapor channels did not result in coherent solutions and was usually subject to higher estimation errors, while the IR channel contributes to the improvement of the visible channel estimation. In the energy, the parameters C and λ define the weights of the smoothing term and the amount of anisotropic behavior of Nagel-Enkelmann operator, respectively. These parameters are normalized and they depend on other two parameters, α and s. In the tests, the best configuration for both sequences is $\alpha = 6.67$ and $s = 0.5$ for multi-channel method and

Table 2

Comparison of the different methods for Vince sequence

Method/Channel	AEE	AAE
SF, VIS 0.8	0.2608	5.2055
SF, WV 6.2	0.5280	12.3891
SF, WV 7.3	0.4992	11.9396
SF, IR 10.8	0.3413	7.8710
MC Avg.	0.1926	3.8432
MC Max.	**0.1895**	**3.7696**

Table 3

Comparison of the different methods for NAtl sequence

Method/Channel	AEE	AAE
SF, VIS 0.8	0.1704	4.4748
SF, WV 6.2	0.5813	15.1845
SF, WV 7.3	0.4776	12.6222
SF, IR 10.8	0.3064	7.7821
MC Avg.	0.1593	4.1831
MC Max.	**0.1350**	**3.4246**

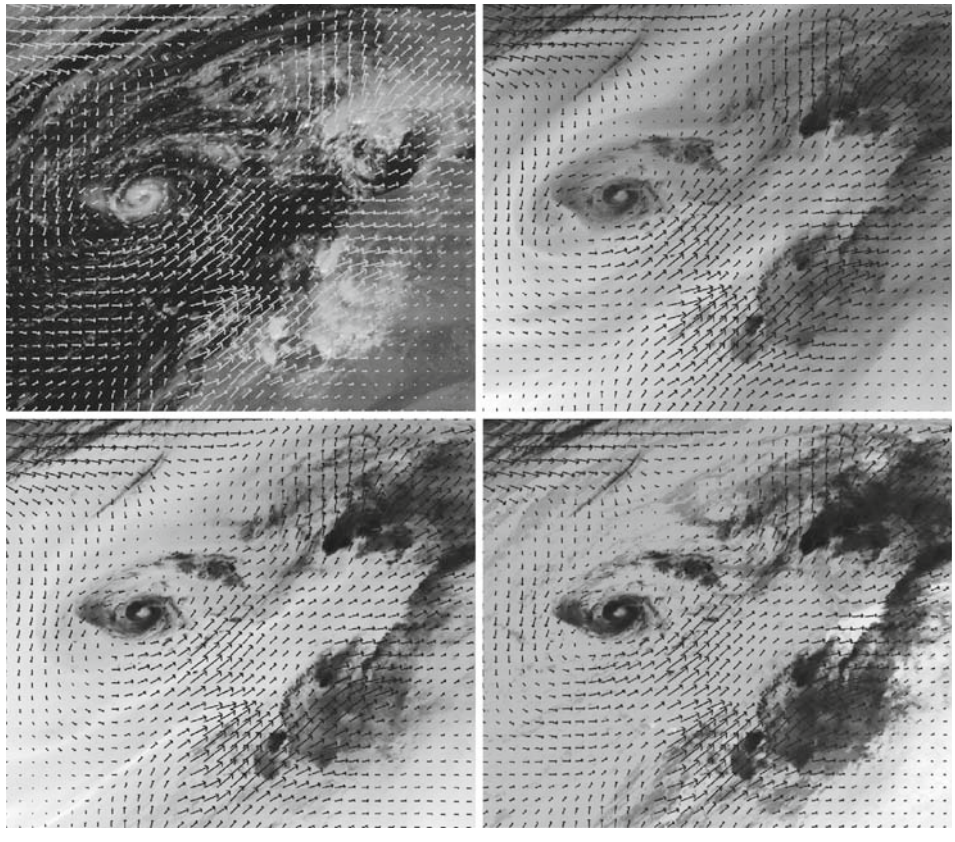


Figure 6
Sequence 1. At the top: the vector field displacement for VIS 0.8 and WV 6.2 channels. At the bottom: The vector field displacement for WV 7.3 and IR 10.8 channels.

$\alpha = 10$ and $s = 0.5$ for single channel method. The α and s are very stable, small changes in their values which do not lead to significant changes of the estimation errors.

Although variational methods compute dense solutions, we display the motion field with arrows every 18 pixels to simplify the visualization.The MSG images we used in the tests are 1024×1024, and cover the area over the North Atlantic (Figure 1). Regions of interest have been selected in each image sequence.

6.1. Sequence 1. North Atlantic (October 8, 2005), Hurricane Vince

Vince was an extraordinary meteorological event, since it was the hurricane farther east than any other currently known, and the first tropical cyclone to reach the Iberian Peninsula. It started as a subtropical storm at 06:00 UTC on October 8, 2005 in the southeast of the Azores Islands, reaching a category 1 hurricane on the Saffir-Simpson Hurricane Scale at 18:00 UTC 9 October 2005, in the northwest of Funchal in the

Madeira Islands. Then, it began to weaken and a few hours later, at 00:00 UTC on October 10, it decayed back to a tropical storm. Vince weakened rapidly during that day as it approached the Iberian Peninsula, where it arrived as a tropical depression during the first hours of October 11 (FRANKLIN, 2006).

In Figure 1 top, we show the images provided by four different channels of the Meteosat Second Generation (MSG) satellite, from Eumetsat, at two different time instants of the Vince sequence. The left column shows the 0.81 μm visible channel, the columns in the middle show, respectively, the 6.25 μm and 7.35 μm water vapor channels and the column on the right the 10.8 μm infrared channel.

In sequence 1, we have selected a region of 642 $\times$ 559 pixels (Fig. 7). In this region we can see the hurricane Vince and a wide cloud over the Northwest coast of Africa. Two kinds of motion are predominant, a rotational by Vince and a translational given by the cloud placed over Africa. In Figure 6, we show the cloud motion estimation of each channel.

As we can see in Figure 7, the influence of hurricane Vince in the region is high. This effect produces that the predominant motion in the region will be the rotation. The power of the winds around the hurricane is high and this phenomenon produces a strong vorticity of the estimated vector field. The second important motion moves upward from Africa. This motion increases in magnitude when it is strengthened by the winds from Vince. The motion obtained by the different IR channels is quite similar. There are differences with respect to the visible channel. In the hurricane sequences a spinning effect of the surrounding clouds appears. This effect is clearer in the visible channel than in the others. For this reason, the contribution of this channel is higher.

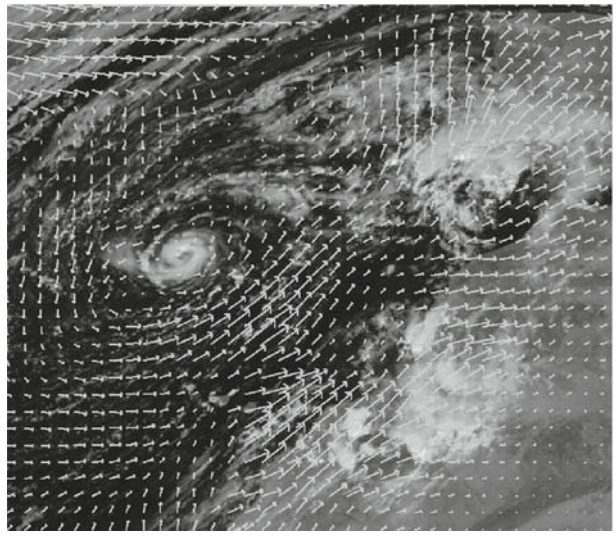


Figure 7
Sequence 1. Flow field obtained with the multi-channel method (Max. gradient strategy).

In Figure 7, we can see the results obtained using the information from the four channels. In Table 2, the multi-channel method offers much better estimations than those of the single channel. The improvement is 27.34% and 27.58% for the AEE and AAE, respectively. There is a little difference (over 1.61% for AEE and 1.92% for AAE) between the two criteria used in the multi-channel method.

In Figure 8, an image of the AEE between the ground truth and the best multi-channel estimation is shown. As we can see, the estimation is good in the whole image except in the boundary and the African coast.

6.2. Sequence 2. North Atlantic (June 5, 2004)

In figure 1 bottom, as was done in the previous case, we present images from the four channels used in our experiments, provided by Eumetsat on June 5, 2004 over the same geographical area. In this sense, we have from left to right the 0.81 μm visible channel, the 6.25 μm and 7.35 μm water vapor channels and the 10.8 μm infrared channel, respectively.

In sequence 2, the size of the selected region is 559 $\times$ 575 pixels (Figure 10). Two main clouds structures are present in the region. The first one is a squall approaching from the Atlantic to the British Islands, placed at the top. The second one is a dense and

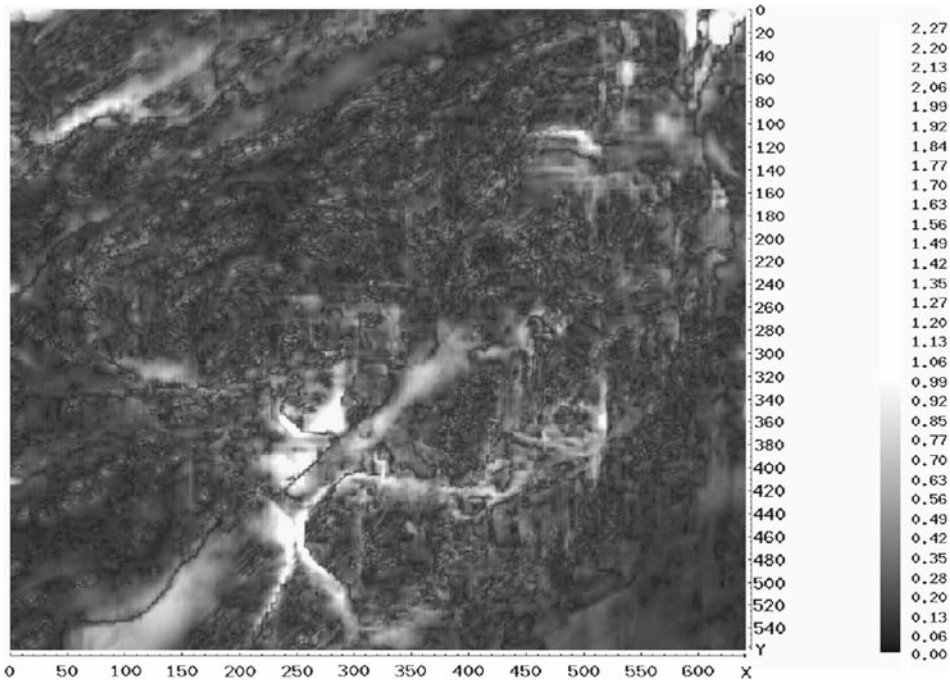


Figure 8
The Average Euclidean Error between the ground truth and the best estimation with the multi-channel method.

wide cloud close to the Iberian peninsula, placed at the bottom. The predominant motion is rotational. In Figure 9, we show the cloud motion estimation of each channel, separately. The solutions obtained by each channel are similar in magnitude. However, the visible channel makes more contribution to the orientation and to the detection of the vorticity. In Figure 9 top, the visible channel is the only one that detects properly the squall vorticity. In the other channels the motion detected is almost translational. A similar effect develops in the cloud structure located in front of the Iberian peninsula. For this reason, the visible channel more prominently higher contributes to the detection of rotational motion. In Figure 10, we can see the result obtained applying the information

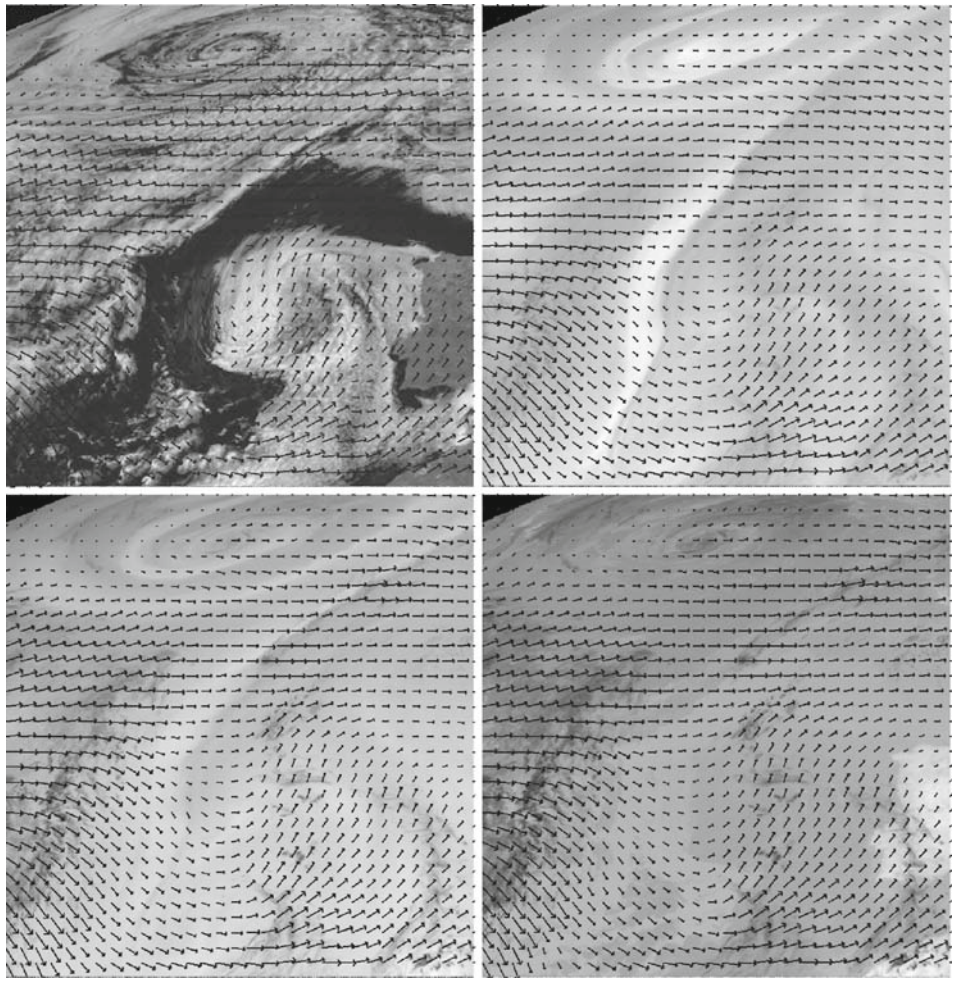


Figure 9
Sequence 2. At the top: the vector field displacement for VIS 0.8 and WV 6.2 channels. At the bottom: the vector field displacement for WV 7.3 and IR 10.8 channels.

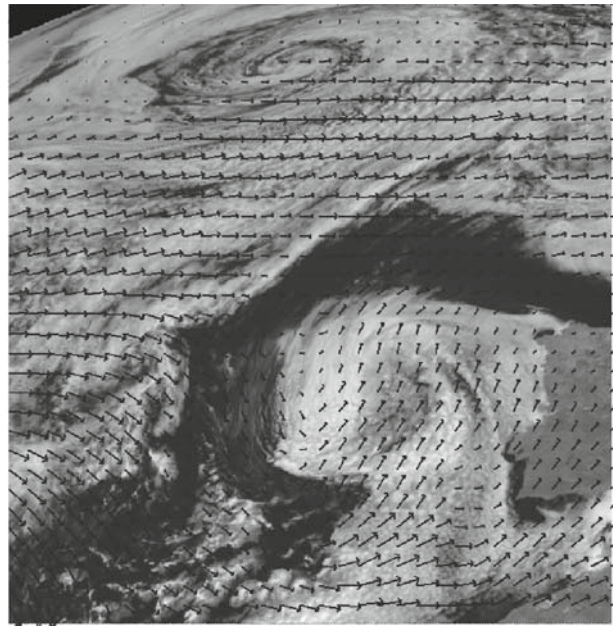


Figure 10
Sequence 2. Motion estimation obtained with the multi-channel method (Max. gradient strategy).

from the four channels using the previously defined weights. Due to the weights, the visible channel information will be predominant, although the rest of the channels contribute to the solution smoothing it.

As we can see in table 3, the best estimation with the multi-channel and single channel method is at least 20.77% better for the AEE. For the AAE the improvement reaches 23.47%. As occurs in the Vince sequence, the differences in the errors between the two strategies used in the multi-channel method are small. Figure 11 depicts an image of the AEE between the ground truth and the best multichannel estimation.

7. Conclusions

Multi-channel meteorological satellite image analysis is a challenging problem. In this paper, we propose use of a variational approach to deal with some of the standard problems in this field.

First, we analyze the problem of cloud structure smoothing and classification: we design several filters, based on variational techniques to smooth the boundary of the provided cloud layer classification regions and also to smooth the different channels. In order to preserve the discontinuities of the channels in the boundary of the classification cloud layers, we perform the smoothing filter separately in each layer. Numerical experiments illustrate the smoothing behavior of the filters.

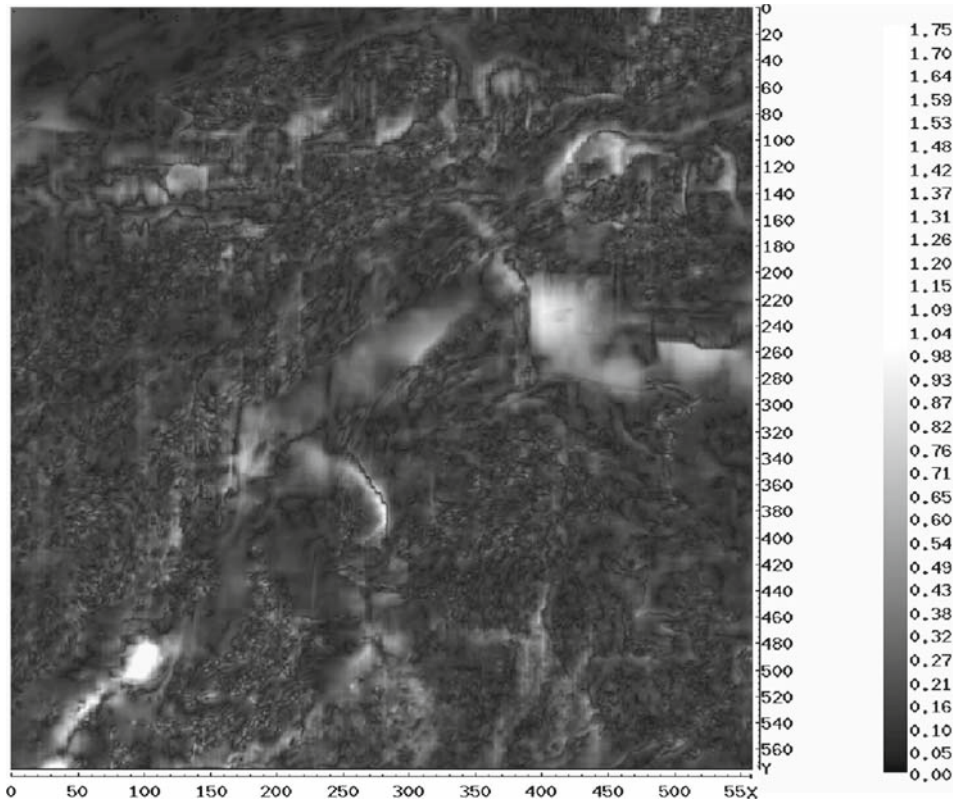


Figure 11
The Average Euclidean Error between the ground truth and the best estimation with the multi-channel method.

Second, in order to estimate the cloud structure motion across the satellite image sequence, we extend a variational motion estimation technique to deal with multi-channel satellite image sequences. The main idea is to combine the information of all channels in order to estimate the cloud structure motion.

To illustrate our experiments, we have developed a freely available software which performs 3D visualization of the cloud structure layers and of the 3D motion vectors. The altitude component of the cloud structure layers is estimated from an estimation of the temperature based on the infrared channel IR 10.8 μm.

In order to perform a quantitative comparison between the different motion estimation methods, we designed a synthetic experiment using a realistic motion model. We ascertained that, using the proposed multi-channel method, the accuracy of the motion estimation is improved significantly as compared to a single channel motion estimation.

Acknowledgements

We would like to acknowledge the Laboratoire de Météorologie Dynamique (CNRS, UMR 8539, France) for providing the experimental data. This work has been funded by the European Commission under the Specific Targeted Research Project FLUID (contract no. FP6-513663).

References

Alvarez, L., Weickert, J., and Sánchez, J (2000), *Reliable estimation of dense optical flow fields with large displacements*, Internat. J. Comp. Vision *39*(1), 41–56.

Alvarez, L., Castano, C., García, M., Krissian, K., Mazorra, L., Salgado, A., and Sánchez, J (2006), *3D atmospheric cloud structures: Processing and visualization*, Video presentation at IEEE Conf. on Computer Vision and Pattern Recognition.

Ameur, Z., Ameur, S., Adane, A., Sauvageot, H., and Bara, K. (2004), *Cloud classification using the textural features of meteosat images*, Int. J. Remote Sensing *25*(10), 4491–4503.

Aubert, G., Deriche, D., and Kornprobst, P. (1999), *Computing optical flow via variational techniques*, SIAM J. Appl. Math. *60*(1), 152–182.

Baker, W. (1991), *Utilization of satellite winds for climate and global change studies*. In *Proc. NOAA Conf. on Operational Satellites: Sentinels for the Monitoring of Climate and Global Change, Global Planetary Change*, vol. 4, pp. 157–163, special issue.

Barron, J., Fleet, D., and Beauchemin, S. (1994), *Performance of optical flow techniques*, IJCV *12* (1), 43–77.

Beauchemin, S., and Barron, J. (1995), *The computation of optical flow*, ACM Comp. Surveys *27*(3), 433–467.

Black, M., and Anandan, P. (1996), *The robust estimation of multiple motions: Parametric and piecewise-smooth fields*, Comp. Vision and Image Understanding *63*(1), 75–104.

Brad, R., and Letia, I. (2002), *Cloud motion detection from infrared satellite images*, In: *Second International Conference on Image and Graphics*, SPIE, vol. 4875, pp. 408–412.

Corpetti, T., Memin, E., and Perez, P. (2002), *Dense estimation of fluid flows*, IEEE Trans on Pattern Analysis and Machine Inteligence *24*(3), 365–380.

Côté, S., and Tatnall, A. (1995), *A neural network-based method for tracking features from satellite sensor images*, Int. J. Remote Sen. *16*(16), 3695–3701.

Deriche, R., Kornprobst, P., and Aubert, G. (1995), *Optical-flow estimation while preserving its discontinuities: A variational approach*, In *ACCV*, pp. 71–80.

Evans, A. (2006), *Cloud motion analysis using multichannel correlation-relaxation labeling*, IEEE Geosci. Remote Sensing Lett. *3*(3), 392–396.

Franklin, J. (2006), *Tropical cyclone report: Hurricane vince*, Draft, National Hurricane Center.

Galvin, B., McCane, B., Novins, K., Mason, D., and Mills, S. (1998), *Recovering motion fields: An evaluation of eight optical flow algorithms*, In British Machine Vision Conference.

Hasler, A. (1990), *Stereoscopic measurements*, In Rao, P., Holms, S., Anderson, R., Winston, J., Lehr, P. (eds.), Weather Satellites:Systems, Data and Environmental Applications, Ameri. Meteor. Soc., Boston, MA.

Héas, P., Mémin, E., Papadakis, N., and Szantai, A. (2006), *Layered estimation of atmospheric mesoscale dynamics from satellite imagery*, IEEE Transact. Geosci. Remote Sensing (under revision).

Horn, B., and Schunck, B. (1980), *Determining optical flow*, MIT Aritificial Intelligence Laboratory.

Inoue, T. (1987), *A cloud type classification with noaa-7 split-window measurements*, J. Geophy. Res. *92*(D4), 3991–4000.

Jähne, B., and Haussecker, H. (2000), *Performance characteristics of low-level motion estimators in spatiotemporal images*, In *Proceedings of the Theoretical Foundations of Computer Vision*, TFCV on Performance Characterization in Computer Vision (Kluwer, B.V., Deventer, The Netherlands) pp. 139–152.

Kambhamettu, C., Palaiappan, K., and Hasler, A. (1995), *Coupled, multi-resolution stereo and motion analysis*, In *Proc. IEEE Int'l Symp. Computer Vision*, pp. 43–48.

LEESE, J., NOVAK, C., and CLARK, B. (1971), *An automated technique for obtaining cloud motion from geosynchronous satellite data using cross correlation*, J. App. Meteor. *10*, 118–132.

MITICHE, A., and BOUTHEMY, P. (1996), *Computation and analysis of image motion: a synopsis of current problems and methods*, Int. J. Comput. Vision *19*(1), 29–55.

MUKHERJEE, D., and ACTON, S. (2002), *Cloud tracking by scale space classification*, IEEE Transact. Geosci. Remote Sensing *40*(2), 405–415.

NAGEL, H.H., and ENKELMANN, W. (1986), *An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences*, IEEE Trans Pat. Anal. and Mach. Intell. *8*, 565–593.

OTTENBACHER, A., TOMASSINI, M., HOLMLUND, K., and SCHMETZ, J. (1997), *Low-level cloud motion winds from meteosat high-resolution visible imagery*, Weather and Forecasting *12*(1), 175–184.

PAPENBERG, N., BRUHN, A., BROX, T., DIDAS, S., and WEICKERT, J. (2006), *Highly accurate optic flow computation with theoretically justified warping*, Internat. J. Comp. Vision *67*(2), 141–158.

PHILLIPS, D., SMITH, E., and SUOMI, V. (1972), *Comment on 'an automatied technique for obtaining cloud motion from geosynchronous satellite data using cross-correlation'*, J. Appl. Meteor. *11*, 752–754.

PROESMANS, M., VAN GOOL, L., PAUWELS, E., and OOSTERLINCK, A. (1994), *Determination of optical flow and its discontinuities using non-linear diffusion*, In ECCV '94, Proc. Third European Conference, vol. II on Computer Vision (Springer-Verlag, London, UK), pp. 295–304.

SCHMETZ, J., HOLMLUND, K., JOFFMAN, J., STRAUSS, B., MASON, B., GAERTNER, V., KOCH, A., and VAN DE BERG, L. (1993), *Operational cloud motion winds from meteosat infrared images*, J. Appl. Meteor. *32*, 1206–1225.

SCHMETZ, J., PILI, P., TJEMKES, S., JUST, D., KERKMANN, J., ROTA, S., and RATIER, A. (2002), *An introduction to meteosat second generation (msg)*, Am. Meteor. Soc. pp. 977–992.

SZANTAI, A., and DÉSALMAND, F. (2005), *Report 1: Basic information on msg images*, Draft, Laboratoire de Météorologie Dynamique.

VELDEN, C., OLANDER, T., and WANZONG, S. (1998), *The impact of multispectral goes-8 wind information on atlantic tropical cyclone track forecasts in 1995. Part i: Dataset methodology, description, and case analysis*, Monthly Weather Rev. *126*(5), 1202–1218.

WEICKERT, J., and SCHNÖRR, C. (2001), *Variational optic flow computation with a spatio-temporal smoothness constraint*, J. Mathe. Imaging and Vision *14*(3), 245–255.

WU, Q. (1995), *A correlation-relaxation-labeling framework for computing optical flow: Template matching from a new perspective*, IEEE Transact. Pattern Analysis and Machine Intelligence *17*(9), 843–853.

YOUNG, G., and CHELLAPPA, R. (1990), *3D motion estimation using a sequence of noisy stereo images: Models, estimation, and uniqueness results*, IEEE Transact. on Pattern Analysis and Machine Intelligence *12*(8), 735–759.

ZHOU, L., KAMBHAMETTU, C.D.B.G., PALANIAPPAN, K., and HASLER, A. (2001), *Tracking nonrigid motion and structure from 2D satellite cloud images without correspondences*, IEEE Transact. Pattern Analysis and Machine Intelligence *23*(11), 1330–1336.

(Received March 30, 2007, revised April 3, 2008, accepted April 3, 2008)
Published Online First: July 11, 2008

Pure appl. geophys. 165 (2008) 1095–1115
0033–4553/08/061095–21
DOI 10.1007/s00024-008-0348-2

Pure and Applied Geophysics

3D Gravity Inversion by Growing Bodies and Shaping Layers at Mt. Vesuvius (Southern Italy)

GIOVANNA BERRINO,[1] and ANTONIO G. CAMACHO[2]

Abstract—To improve our knowledge of the structural pattern of Mt. Vesuvius and its magmatic system, which represents one of the three volcanoes located in the Neapolitan area (together with Campi Flegrei and Ischia; southern Italy), we analyze here the Bouguer gravity map that is already available through its interpretation by means of 2.5-dimensional modelling. We have carried out a three-dimensional interpretation using a new and original algorithm, known as 'Layers', that has been especially processed for this purpose. Layers works in an automatic and non-subjective way, and allows the definition of the structural settings in terms of several layers, each representing a specific geological formation. The same data are also interpreted in terms of isolated and shallow anomalous density bodies using a well tested algorithm known as 'Growth'. We focus our inversions on the Mt. Vesuvius volcano, while globally analyzing the entire Neapolitan area, in order to investigate the deep structures, and in particular the deep extended 'sill' that has been revealed by seismic tomography.

The final models generally confirm the global setting of the area as outlined by previous investigations, mainly for the shape and depth of the carbonate basement below Mt. Vesuvius. The presence of lateral density contrasts inside the volcano edifice is also shown, which was only hypothesized in the 2.5-dimensional inversion. Moreover, the models allow us to note a high density body that rises from the top of the carbonate basement and further elongates above sea level. This probably represents an uprising of the same basement, which is just below the volcano and which coincides with the V_P and V_P/V_S anomalies detected under the crater. The three-dimensional results also reveal that the two inversion methods provide very similar models, where the high density isolated body in the Growth model can be associated with the rising high density anomaly in the Layers model. Taking into account the density of these modelled bodies, we would also suggest that they represent solidified magma bodies, as suggested by other studies. Finally, we did not clearly detect any deep anomalous body that can be associated with the sill that was suggested by seismic tomography.

Key words: Gravity, Bouguer anomaly, Mt. Vesuvius, three-dimensional inversion, model exploration, algorithms.

1. Introduction

As already well known, the gravity method is a powerful tool for the exploration of the subsoil. It is largely applied to understand volcanic activity too, and it can also be

[1] Istituto Nazionale di Geofisica e Vulcanologia, sezione di Napoli 'Osservatorio Vesuviano', Via Diocleziano 328, 80124 Napoli, Italy. E-mail: berrino@ov.ingv.it

[2] Instituto de Astronomía y Geodesia (CSIC-UCM), Plaza Ciencias 3, 28040 Madrid, Spain.

applied to volcanic areas where knowledge of the structural setting is helpful for the outlining of routes of probable magma uprisings. This is the case for Mt. Vesuvius.

Together with Campi Flegrei and Ischia, Mt. Vesuvius is one of the three active volcanoes that are located in the Neapolitan area (southern Italy; Fig. 1). These Neapolitan volcanoes lie within the Campanian Plain, a graben that is bordered by a Mesozoic carbonate platform that stretches from Mt. Massico, deepens to more than 3 km in its central part, and then re-emerges in the Sorrento Peninsula. The Campanian Plain is bordered on the NE by NW-SE-trending faults, and on the S and the N by a horst that is limited by NE-SW trending faults. This important graben is filled with volcanic deposits and continental and marine clastic deposits (BALDUCCI *et al*., 1985). The buried geometry of the carbonatic basement has been outlined through gravity data on land (OLIVERI DEL CASTILLO, 1966; CARRARA *et al*., 1973; CAMELI *et al*., 1975; LUONGO *et al*.,

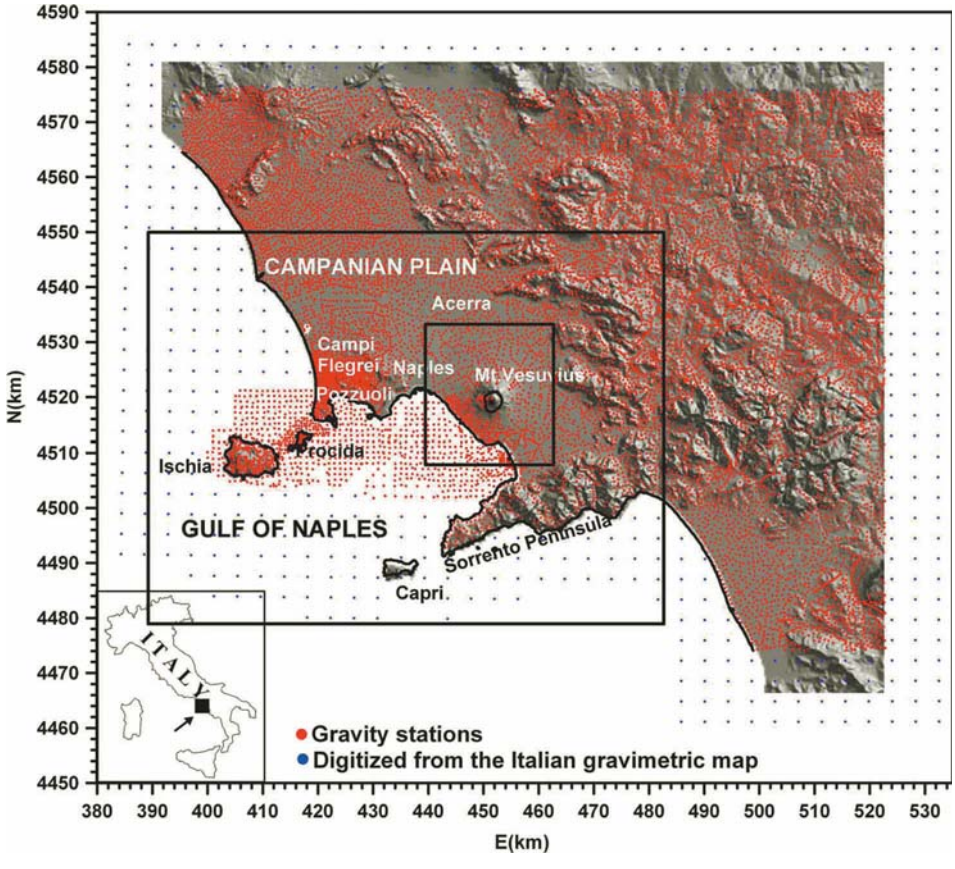


Figure 1
Distribution of the gravity stations (red points). The blue points indicate the values digitized on the Italian Gravity Map (CARROZZO *et al*., 1986; see text for details). The two black rectangles show the two 3D inversion areas (see text for details).

1988; FERRI *et al*., 1990; CUBELLIS *et al*., 1995). Many studies have suggested that the main feature of the central part of the Campanian Plain is a structural depression, known as the 'Acerra depression' (BARBERI *et al*., 1978; SANTACROCE, 1987). Another important depression, known as the 'Pompei graben', was detected by CASSANO and LA TORRE (1987).

Mt. Vesuvius is a strato-volcano that lies about 15 km southeast of Naples, and it consists of an older structure (Mt. Somma) with a nested younger structure (Mt. Vesuvius). It is located on a NE-SW trending fault that borders the southeastern edge of the Acerra depression (MARZOCCHI *et al*., 1993), and it lies on a sedimentary basement. Gravimetry on land (CASSANO and LA TORRE, 1987) and seismic profiles at sea (FINETTI and MORELLI, 1974; FINETTI and DEL BEN, 1986) have shown that this fault displaced the more recent formations. Information about the sedimentary basement below Mt. Vesuvius has been provided by the deep geothermal Trecase well that intercepts the limestone layer at about 1,700 m below sea level (b.s.l.) (BALDUCCI *et al*., 1985). Moreover, the inversion of both on-land and off-shore gravity data has suggested that the sedimentary basement is 11 km thick, with the top at a depth of 2 km (BERRINO *et al*., 1998). This was confirmed by a joint seismic tomographic inversion of first P-wave arrivals along several profiles intersecting the crater, and from gravity data that provided a clear image of the continuous structure of the Mesozoic carbonate basement top as well as of a conduit structure 5 km wide that extends from the surface to the maximum depth of the model (6 km) (TONDI and DE FRANCO, 2003, 2006). No significant evidence has been seen for the existence of a shallow magma chamber embedded in the basement (ROSI *et al*., 1987; CORTINI and SCANDONE, 1982).

A recent seismic tomography study was carried out to define the evidence within the Vesuvius magmatic system of an extended (at least 400 km^2) low-velocity layer at about 8 km in depth, which would represent an extended sill with magma interspersed in a solid matrix (AUGER *et al*., 2001). This body was also modelled by new isotopic data (CIVETTA *et al*., 2004) and by a new inversion of P-wave and S-wave arrival times for local earthquakes, highlighting a lower V_s velocity below the Mt. Vesuvius cone in a 0.35-km-thick layer (NUNZIATA *et al*., 2006). Moreover, a joint inversion of P-wave and S-wave arrival times (from local earthquakes) and shot data collected during the TOMOVES 1994 and 1996 experiments showed the presence of a high V_p and V_p/V_s anomaly that is located around the crater axis, between 0 km and 5 km in depth, which involves the volcano edifice and the carbonate basement. This anomaly has been interpreted in terms of magma quenching along the main conduit, because of the exsolution of magmatic volatiles (DE NATALE *et al*., 2004).

To improve our knowledge of the structural pattern of Mt. Vesuvius and its magmatic system, we analyze here the Bouguer gravity map that is already available through its interpretion by means of 2.5-dimensional (2.5-D) modelling (BERRINO *et al*., 1998). We have carried out a 3D interpretation using a new and original algorithm that was specifically realized for this study and that starts from a known algorithm (CAMACHO *et al*., 2000, 2002). We have focused our inversions on the Mt.Vesuvius volcano,

although we also globally analyze the entire Neapolitan area to investigate the deep structures, and particularly that known as the 'sill' that was revealed by a seismic tomography study carried out by AUGER *et al*. (2001).

A description of this first version of the new algorithm is given, together with the results obtained.

2. *Gravity Data and Previous Interpretations*

Here we use a Bouguer anomaly map that consists of both on-land and off-shore gravity data. To complete the existing gravity map that was limited to on-land data, a sea-gravity survey was carried out in the Gulf of Naples during five cruises that lasted from 1988 to 1994 (BERRINO *et al*., 1991, 1998), whereby 850 off-shore points where measured. In this way the Bouguer map was created, and it provides a global set of 2,876 gravity values (BERRINO *et al*., 1998, 2008). All of the data have been made uniform and globally re-analyzed through the referencing of the gravity values to a new absolute gravity station set up in Naples in 1986 (BERRINO, 1995), which also belongs to the new Italian **'**Zero Order**'** Gravity Net (BERRINO et al., 1995). Later, in the framework of a cooperation in the TOMOVES Project (ACHAUER *et al*., 1999, 2000), additional gravity data became available for the whole Campanian Plain (provided by P. Capuano). These data were collected and combined with the previous dataset, such that after an additional revision and data cleaning (SCALA, 2002), a new gravity dataset of 17,225 gravity values was obtained, as shown as red points in Figure 1. In this case too the updated dataset was recomputed and linked to the absolute gravity station in Naples. The detailed references relating to the available on-land gravity data and the information as to how the offshore gravity data were collected, integrated with the on-land data and globally analyzed, as well as the information about the interpretation of some of the previous geophysical investigations, are all given in BERRINO *et al*. (1998, 2008).

Two Bouguer gravity maps were obtained with reference to the 1980 Ellipsoid (MORITZ, 1984), using the density values of 2,200 kg/m^3 and 2,400 kg/m^3, to calculate the Bouguer and terrain effects. The first of these values is more suitable for volcanic areas (BERRINO *et al*., 1998), while 2,400 kg/m^3 is more appropriate for the global interpretation of the whole Campanian Plain (SCALA, 2002). Moreover, this second case allowed the area investigated to be enlarged through the addition of a border about 10 km larger, which was obtained by digitizing 517 anomaly values (Fig. 1, blue points) from the Italian Bouguer gravity map (CARROZZO et al., 1986).

Detailed descriptions of the Bouguer gravity maps are given in BERRINO *et al*. (1998, 2008) and SCALA (2002). However, their main features can be summarized as follows: A strip of maximum gradient runs almost parallel to the Sorrento peninsula, turns towards the southwest at the southern sector of Somma-Vesuvius, and ends in the southern part of the Gulf of Pozzuoli, in a broad gravity minimum. The Somma-Vesuvius and Campi Flegrei volcanoes are settled at the southern edge of the large gravity minimum, to the

north of Naples. A well defined gravity minimum that spans Campi Flegrei is evident. Strong gradients border the Island of Ischia. The Vesuvian area is characterized by a Bouguer anomaly of small extension and amplitude, which follows a tortuous pattern due to the presence of local minima and maxima. A vast gravity low southeast of Somma-Vesuvius is the main feature in this area; it corresponds to the so called 'Pompei graben' (CASSANO and LA TORRE, 1987).

The Bouguer anomaly gravity map has already been interpreted by means of 2.5D modelling (WON and BEVIS, 1987; FEDI, 1988) along a series of profiles that have provided information about the main volcanic structures, and particularly about the shape and depth of the limestone and crystalline basements (BERRINO *et al*., 1998, 2008; SCALA, 2002). This kind of inversion has also allowed the building up of a pseudo-3D pattern of the limestone basement and the delineation of the main tectonic structures under the Neapolitan volcanoes (BERRINO *et al*., 1998, 2008).

In this way, no important information has been obtained about very shallow and isolated bodies, and therefore a 3D interpretation using a new and original algorithm has been used here, with the aim of better defining the shallow density distribution. We have used the Bouguer anomaly map that has been reduced with the density value of 2,200 kg/m^3 because we have limited our first investigation to the volcanic area. Therefore, we have selected an area of about 20 km $\times$ 20 km (see Figure 1, small black rectangle) centered on Mt. Vesuvius, where 616 gravity stations lie. First, we focused our inversions on a local and shallow body below the volcano, then we attempted to globally analyze the entire Neapolitan area in order to investigate deep structures, and mainly the sill that was revealed by the seismic tomography study. This second aspect was carried out by analyzing the 6,203 gravity values inside the area (90 km $\times$ 70 km) that is illustrated with the larger black rectangle in Figure 1.

Although we aimed to design objective models without any subjective preliminary information or input model, the information provided by the previous 2.5D interpretation has been taken into account as a reference, if necessary and when possible.

The Bouguer anomaly map that has been reduced with a 2,200 kg/m^3 reference density and that spans the larger area of analysis is shown in Figure 2.

3. The 3D Gravity Inversion Method

3.1. Introduction to the Gravity Method

Very extensive results have been obtained for gravity modelling by methods of 'trial and error'. For instance, the IGMAS method is an interactive, graphical computer system for the interpretation of potential fields (gravity and magnetic) by means of numerical simulations (GÖTZE and LAHMEYER, 1988). These direct methods are based on strong personal experience and *a priori* knowledge of the structure at depth, and they have a more or less subjective character.

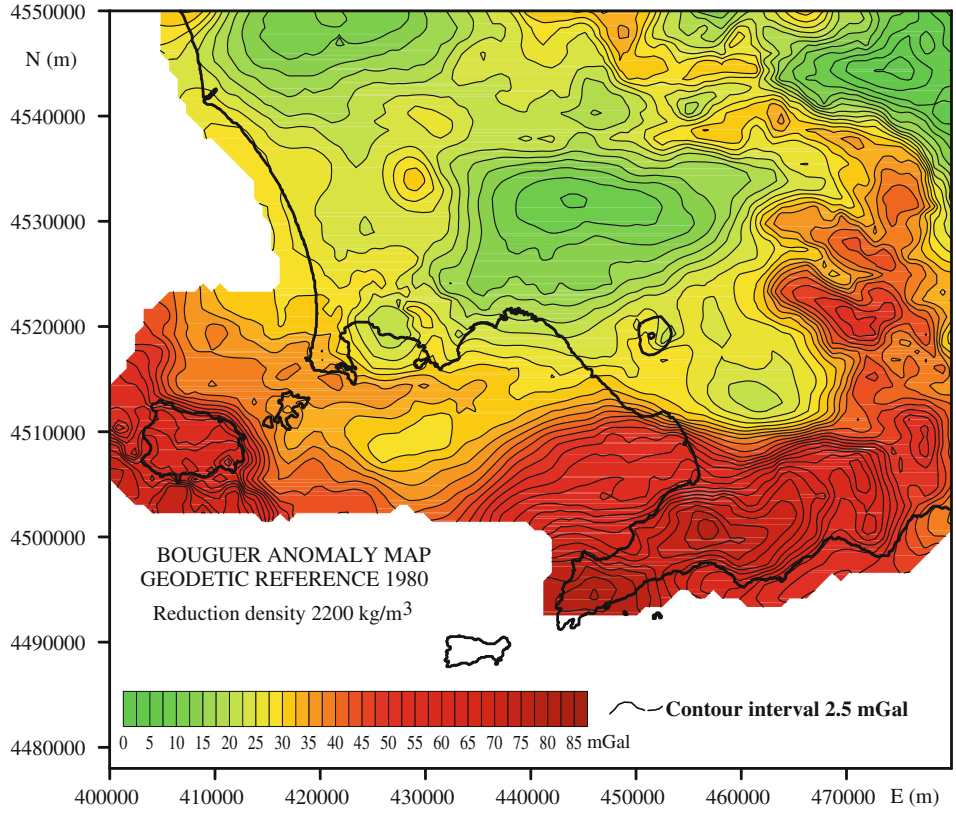


Figure 2
Bouguer anomaly map, reduced with a 2,200 kg/m^3 reference density, for the large selected area (see Fig. 1 and text for details).

We look for the determination, in a non-subjective way, of a model of the subsoil density distribution that can reproduce the observed gravity anomaly. Taking into account the information coming from other geophysical investigations (mostly seismic information), we wanted to describe the 3D anomalous density structures mainly by means of several sub-horizontal discontinuity layers (density discontinuities) with irregular boundaries. This is a traditional use of gravity inversion (e.g., studies of sedimentary basins). Considering only one irregular discontinuity surface, the inversion process is not hard to detect, and many studies that have been providing suitable procedures to obtain inversions for just one discontinuity surface are available (e.g., Radhakrishna Murthy and Jagannadha Rao, 1989; Rama Rao *et al*., 1999; Gallardo-Delgado *et al*., 2003). A more problematic question is to obtain a non-subjective inversion model when several discontinuity layers are simultaneously considered. In this case, the assignment of the anomalous density structures among the several layers is not so easy to determine. In general terms, short wavelength features of the gravity anomaly should be assigned to

shallow surfaces, and long wavelength features should be mostly assigned to deep discontinuity surfaces. In this sense, several studies have addressed the inversion process in a frequency domain (e.g., CHAKRABORTY and ARGAWAL, 1992), similar to studies in magnetic prospecting.

Clearly, short wavelength features will correspond to shallow structures; long wavelength features can instead be associated to deep structures, although they can be also associated to extensive enough and not very deep structures. The classical non-uniqueness problem in potential fields requires some additional constraints, which will possibly come from good geological and/or geophysical data, or from general mathematical hypotheses.

On the other hand, gravity inversion by means of the adjustment of sub-horizontal surfaces falls into the nonlinear inversion problems. This requires using some iterative or exploratory approaches to obtain a solution.

To face these problems, we adapted the basic ideas of a previous method for 3D inversion (CAMACHO *et al*., 2002), which we modified for use in our new context. In the previous method (the Growth method), the anomalous model is described as being composed of isolated anomalous bodies, which are constructed in a very free growth process as 3D aggregations of cells (see Fig. 3, left). This method is very interesting for gravity anomalies due to isolated bodies. In a versatile and non-subjective form, and with few constraints, the process can produce 3D models of the anomalous structure (position, depth, size, shape), which are more valuable if suitable values for the density contrast are previously defined. Conversely, some application problems can arise when the causative structure cannot be clearly associated with isolated bodies. In this case, the inversion model will provide a simplified, rather indicative, solution to the inversion problem, that needs further analysis to reach any realistic conclusions. This is the case, for instance, of anomalies due to small distortions of sub-horizontal layers in the subsoil.

Now we want to describe the subsoil model as sub-horizontal layers, where the irregular discontinuity surfaces are constructed by displacing, step by step (according to a system of connected cells, in a growth process), the original flat mean surface (Fig. 3, right). Then, working with similar minimization conditions and similar constructive processes, such as the aggregation of small filled cells in an explorative process of growth, the methodology is modified to allow the aggregation of filled cells that are only connected (up and down) to previous discontinuity surfaces. For a continuous structure (a stratified structure), this looks more realistic than has been described by isolated bodies.

3.2. Inversion Method

The inverse gravimetric problem, namely the determination of a subsurface mass density distribution corresponding to an observed gravity anomaly, has an intrinsic non-uniqueness in its solution (e.g., AL-CHALABI, 1971). Moreover, the data must be considered as insufficient and inaccurate. Nevertheless, particular solutions can be obtained by including additional information about the model parameters (e.g.,

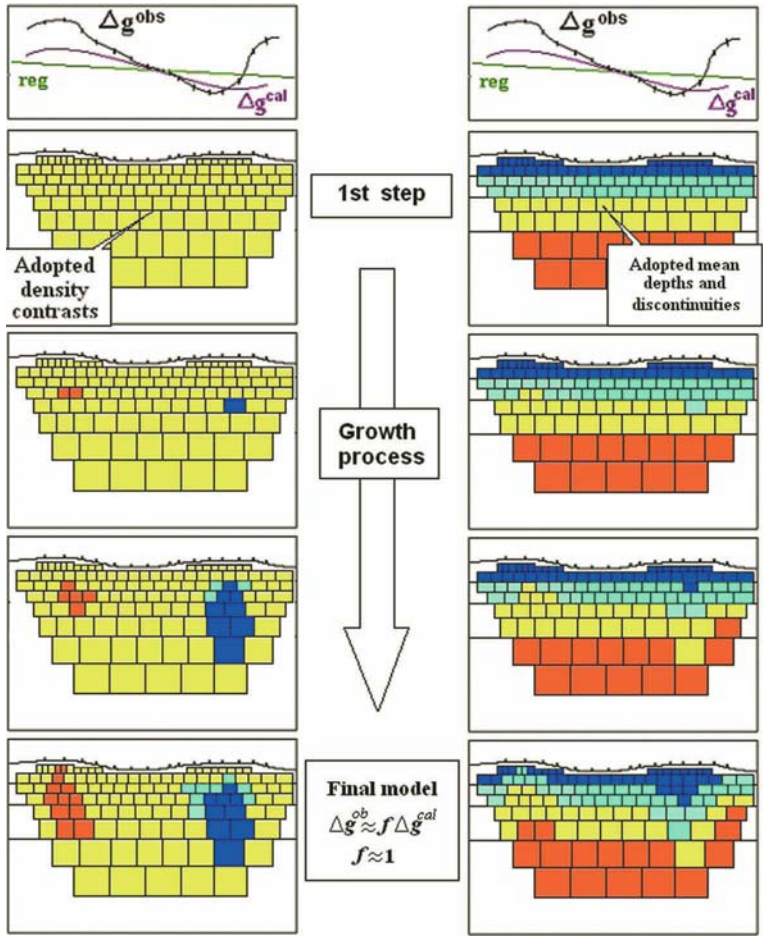


Figure 3

The step by step process in the design of the inversion models: As cells for the Growth model (left) and stratified for the Layers model (right), for the distribution of density according to isolated bodies and sub-horizontal layers, respectively. The geometry of the closed and isolated bodies, together with the discontinuity layers, generates anomalous high and low density areas that are responsible for the anomaly seen.

subsurface structure) and about the data parameters (the statistical properties of inexact data; e.g., a Gaussian distribution). The inversion methods looking for the geometrical properties of anomalous bodies with prescribed density contrast (e.g., PEDERSEN, 1979; BARBOSA *et al*., 1997) correspond to a nonlinear context and offer interesting results that are limited by the validity of the hypothesis used. Unfortunately, linearized techniques depend strongly on the accuracy of the initial estimates of the model parameters (ROTHMAN, 1985). For a fully nonlinear treatment, the methods of random space model exploration often provide the best option (TARANTOLA, 1988; SILVA and HOHMANN, 1983).

Here, we develop a nonlinear inversion method for the geometrical description of the anomalous density structure as sub-horizontal layers. We have named it '*Layers*'.

This method starts from several horizontal layers (up to four in the initial version), which can be introduced *ad hoc*, or conversely, they can be automatically selected in an optimizing approach. Also, several corresponding density contrasts are previously selected (or automatically chosen in a relatively optimizing process). Then, the algorithm works according to a nonlinear explorative approach, to 'deform', or better, to 'shape', the layers step by step, to finally obtain some irregular shapes that can fit the observed anomaly satisfactorily (see Fig. 3, right).

A general tool to describe the geometry of the anomalous mass structure corresponding to irregularities in the sub-horizontal layers is obtained through an aggregation of small parallelepiped cells filled with anomalous mass close to the adopted layers. This procedure can be used to describe general 3D models, as with CAMACHO *et al*. (2000, 2002), although generates a very large number of degrees of freedom for the model. Therefore, a general exploratory inversion approach would be ineffective. An interesting idea was proposed by RENÉ (1986): He applied an exploratory method (in a more restrictive context) not to the global model, but just to every step of its growth process. Under these conditions, the number of degrees of freedom is drastically reduced for each step of the model growth, consequently so the exploratory process becomes very effective.

Let us consider n gravity stations P_i (x_i,y_i,z_i), $i = 1, \ldots, n$ that are not necessarily gridded, which are located on a rugged topography and which have observed anomalous gravity values Δg_i^{obs} (Bouguer anomaly). We assume a mostly Gaussian distribution for the observation uncertainties given by a covariance matrix $\boldsymbol{Q_D}$ (as deduced from analysis of the data). Let us also consider nh horizontal surfaces with depths d_k and density discontinuities $\Delta\rho_k$ (positive differences between the upper and lower media limiting with this surface), for $k = 1, \ldots, nh$.

Our goal is to construct a 3D model that is described as sub-horizontal layers for prescribed mean depths and density discontinuities, and which is 'responsible' for the anomaly observed. As previously indicated, the subsurface volume close to the survey area is dismantled into a global discrete 3D partition of m prismatic elements. The desired solution will be described as an aggregation of some of the prismatic cells filled with prescribed density contrast close to the discontinuity surfaces, thus giving rise to 'shaped' layers. When the filled cell is just below the discontinuity surface, this means that there is an intrusion of low density from the upper medium into the lower one. Conversely, when the filled cell is close and above the discontinuity surface, this means there is an intrusion of high density from the lower medium into the upper one (see Fig. 3, right).

The gravity attraction A_{ij} at the i-th station $P_i(x_i,y_i,z_i)$ due to the j-th prism, per unit density, can be found in PICK *et al*. (1973). Matrix A, with components A_{ij} is the design matrix of the physical configuration problem and includes the effects of rugged terrain, station distribution, subsoil partition, etc. Now the calculated anomaly values for the resulting model are:

$$\Delta g_i^{cal} = \sum_{j \in J_+} A_{ij} \Delta \rho_j - \sum_{j \in J_-} A_{ij} \Delta \rho_j + \Delta g_{reg}, \quad i = 1, ..., N, \tag{1}$$

where J_+ , J_- are the sets of indices that correspond to the cells that are filled and are located up and down, respectively, with respect to the corresponding discontinuity surfaces; Δg_{reg} is a regional smooth trend to be simultaneously adjusted; and J_+ , J_- , Δg_{reg} are the main unknowns to be determined in this inversion process. For the sake of simplicity, we are going to adopt a linear expression for the trend:

$$\Delta g_{reg} = p_0 + p_x(x_i - x_M) + p_y(y_i - y_M), \quad i = 1, ..., N \tag{2}$$

where x_M, y_M are the coordinates of an arbitrary central point for the survey; p_0, p_x, p_y are three unknown values which fit a trend (a 1-degree polynomial surface, simplifying the subsequent formulation).

To solve the inherent non-uniqueness problem, an additional condition of minimization of the model variation can be adopted. Thus, the solution is obtained through a mixed condition between the gravity l_2-fitness and the whole anomalous mass quantity, using a λ parameter for the suitable balance:

$$v^T \boldsymbol{Q_D}^{-1} \boldsymbol{v} + \lambda \, m^T \, \boldsymbol{Q_M}^{-1} \, \boldsymbol{m} = E = \min, \tag{3}$$

where $\boldsymbol{m} = \left(\Delta \rho_{1_1}, ..., \Delta \rho_m\right)^T$ is the anomalous density vector for the m cells of the subsoil partition ($-\Delta \rho_j$ or $\Delta \rho_j$ for the filled cells and zero for those not filled); λ is a positive factor that is empirically fixed and provides the balance between model fitness and anomalous model magnitude (and complexity); $\boldsymbol{Q_D}$ is the covariance matrix (usually a diagonal matrix) that corresponds to the estimated (Gaussian) inaccuracies of the gravity data; and $\boldsymbol{Q_M}$ is a diagonal normalizing matrix whose non-null elements that are the same as the diagonal elements of $\boldsymbol{A^T Q_D^{-I} A}$. The first addend of the minimization functional (3) corresponds to the fit residues weighted with the data quality matrix. The second addend is a weighted addition of the model densities. Nevertheless, taking into account that the covariance matrix $\boldsymbol{Q_M}$ contains the prism volumes as a factor, this second addend is connected with the anomalous mass or magnitude of the model. $v = (v_1,...,v_N)^T$ (T for transpose) is the vector for the gravity residuals for N stations. These are defined as:

$$v_i = \Delta g_i^{ob} - f \Delta g_i^{cal}, \tag{4}$$

where f is a scale factor that allows the fitting of the calculated anomaly for a developing model with respect to the observed values.

The λ parameter governs the application of the minimization conditions with respect to the balance between the total anomalous mass and the residual values. For low λ values, a good fit is obtained, although the anomalous mass may increase excessively and includes some fictitious structures. Conversely, for high λ values, the adjusted model can be too slight, and a poor gravity fit is obtained.

Thus, the inversion process seeks to determine a geometrically anomalous model that is described as an aggregation of filled cells that is connected to discontinuity surfaces and that verifies the minimization condition (3). As previously indicated, we have addressed this nonlinear problem with a process that explores the model possibilities. The exploration of the possibilities for the entire model is substituted by the exploration of several possibilities of growth (cell by cell) for each step of the surface deformation. Thus, the prismatic cells that are connected (up or down) to discontinuity surfaces are systematically tested, step by step.

For one step, some cells have been previously filled to modify the geometry of the initial discontinuity surfaces, although not enough to reproduce the anomaly observed. Now every cell (up and down) that is connected to every actual discontinuity surface is tested. For each cell considered, the linear least-squares problem connected to (3) is solved for unknown parameters f, p_0, p_x, p_y; then the value E of expression (3) is obtained. Once a certain number of cells are randomly selected and verified, we chose to use the smallest value E as the best option for incorporation into the growth approach for the discontinuity geometry.

This process is repeated successively, including a detection of outliers (ROUSSEEUW and LEROY, 1987). In the subsequent steps, the scale value f decreases, and the trend parameters p_0, p_x, p_y reach nearly stable values. The process will stop when f approaches 1, which is reached for a final geometry of the discontinuity surfaces and a final regional trend.

Finally, the solution appears as a 3D distribution of prismatic cells filled with some of the prescribed contrast densities. These prismatic cells are ordered to produce a model of stratified density according to some distribution of the layers limited by sub-horizontal discontinuity surfaces, as illustrated in Figure 3 (right). Moreover, a regional trend supplementing this anomalous mass distribution is also obtained.

This inversion approach has been attempted for the gravity anomaly of the Mt. Vesuvius area, to investigate the structural results that can be obtained in a non-subjective 3D inversion process. As standard for this kind of automatic modelling without a prior hypothesis, the results are valuable as non-subjective information. Of course, some further interpretative subjective work based on some of the initial geological/ geophysical constraints is necessary to 'translate' the model into a more realistic structure.

4. Analyses, Results and Discussion

First of all, from the total of 616 points, for homogeneity purposes we selected 400 points with mutual distances greater than 350 m.

To investigate the shallow structures and possibly the magma system of Mt. Vesuvius, at first the well tested Growth process was adopted. In this case, we did not use any initial constraints and the models obtained are totally non-subjective. Later, taking into account the information from previous geological and geophysical studies

about several sub-soil layers with different lithologies, the Layers process was chosen. We selected the following values for the initial depths and density discontinuities: From a topographical level and $\Delta\rho = 0$ kg/m^3 (very shallow deposits with a density range from 2,000 kg/m^3 to 2,200 kg/m^3); 1,100 m b.s.l. and $\Delta\rho = +200$ kg/m^3 (initial depth for the top of the deeper volcano-sedimentary filling of the Campanian Plain—$\rho = 2{,}400$ kg/m^3); 2,700 m b.s.l. and $\Delta\rho = +200$ kg/m^3 (initial depth for the top of the carbonate formation - $\rho = 2{,}600$ kg/m^3); 9,900 m b.s.l. and $\Delta\rho = +200$ kg/m^3 (initial depth for the top of the crystalline formation - $\rho = 2{,}800$ kg/m^3). Starting with these values, a 3D model was obtained in a nearly automatic process. The resulting residual values show a standard deviation of 363 μGal for the Growth inversion, and 385 μGal for that of the Layers; both show a pattern of non-autocorrelated noise (Figs. 4a, b, respectively). The regional trend (Fig. 5a) obtained simultaneously in the inversion process is characterized by a gravity increase of 1.312 mGal/km towards N170°E. In Figure 5b, the consequent local anomaly is shown: It is generally positive, with very small values of the order of 3–5 mGal around Mt. Vesuvius, and with a very short wavelength, which is indicative of shallow and small isolated bodies. A negative anomaly is seen at the W, S and E sides of the base of the volcanic structure. The positions of the 400 points selected are also shown in Figure 5.

Finally, the resulting 3D models of the isolated bodies and sub-horizontal layers are shown in Figures 6 and 7, respectively, by means of vertical *versus* depth profiles and horizontal deep cross sections.

Figure 6 shows several horizontal cross sections from depths of 0 m to 5,000 m, as six W-E, one N-S, one SE-NW and one SW-NE vertical sections. The vertical section *c*, together with the N-S, SE-NW and SW-NE vertical sections, crosses the Vesuvius crater. All of the profiles reach a depth of 6 km. They all indicate the presence of closed positive and negative density bodies limited to a depth of about 3 km, which are more easily detectable in the horizontal sections. The most significant anomalous density bodies are:

— A positive density body located beneath the crater, that extends towards the NE, where it reaches its maximum depth. It is clearly visible in the *a*, *b* and *c* profiles, in the N-S and mainly in the SW-NE vertical sections;
— a very shallow (from above sea level to some hundreds of meters b.s.l.) negative density body inside the volcano edifice that is clearly visible in the *c* profile;
— several negative density bodies around Mt. Vesuvius, which are mainly W and SE and which were also seen by Tondi and de Franco (2006).

The first two density bodies have already been indicated in these same positions, by Berrino *et al*. (1998), along a 2.5D interpretive profile that coincides with our *c* vertical section; they associated the denser body to lavas and hypothesized a shallow structure that is characterized by density contrast just beneath the volcano edifice. However, they also stressed that the density and the geometry of the bodies inside the volcano were chosen only to obtain an acceptable fitting of the anomaly seen. Here, we highlight that the algorithm produces this model in a very free, automatic and non-subjective way.

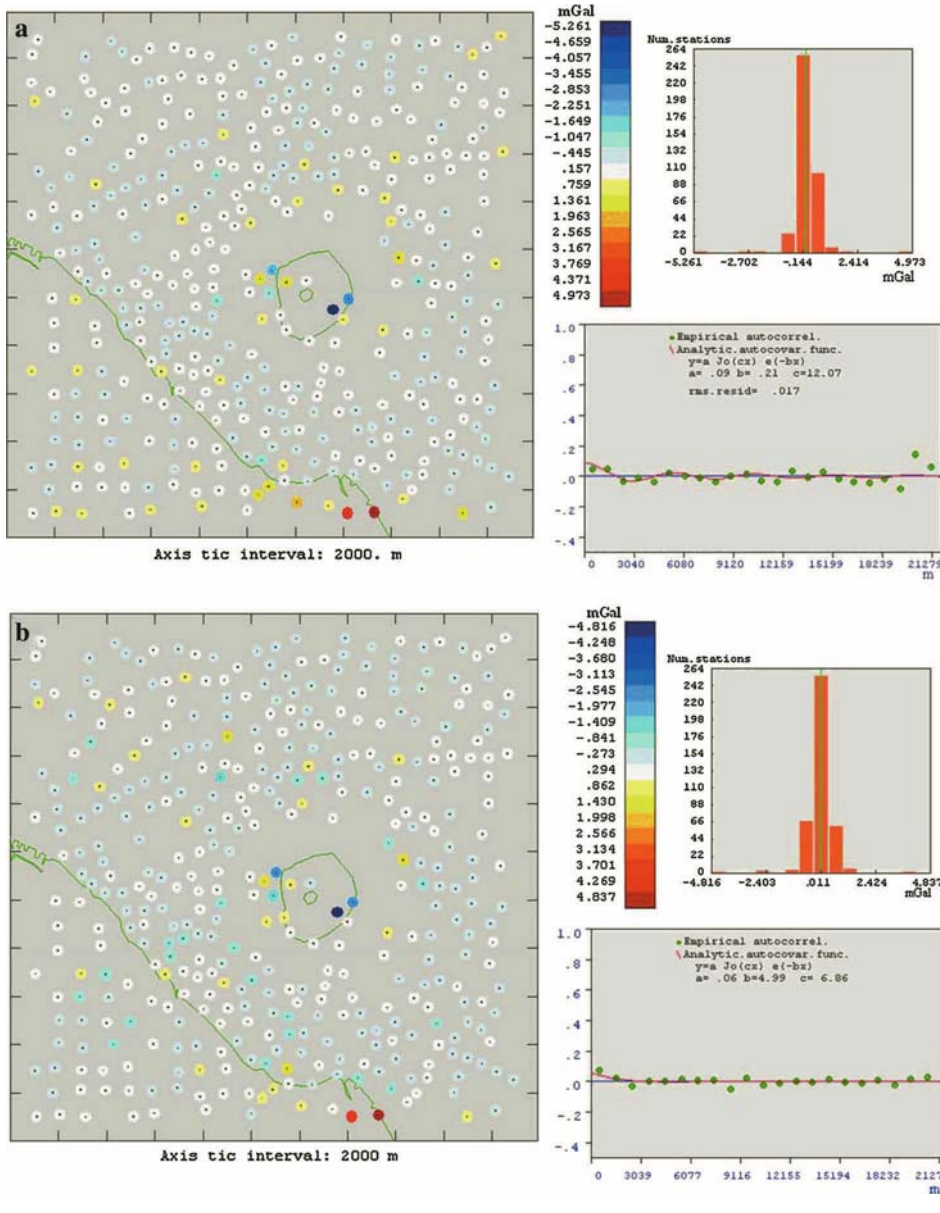


Figure 4
Inversion residuals for (a) the Growth and (b) the Layers models, each including planar distributions, histograms and autocorrelation analyses.

Moreover, high velocity lateral contrast has already been indicated by 2D seismic tomography (Zollo *et al.*, 1996), as well as by a more recent study by De Natale *et al.* (2004) on the inversion of several seismic signals.

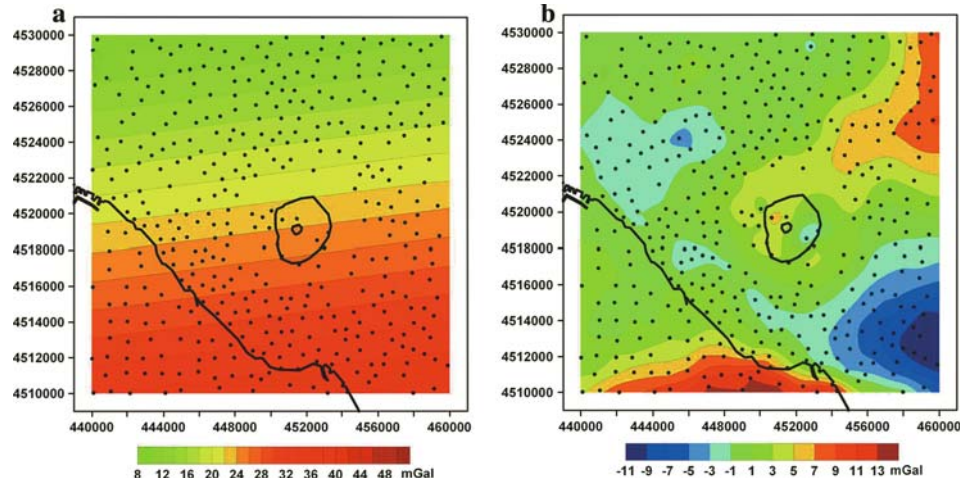


Figure 5
(a) Regional and (b) local gravity anomalies in the small selected area (see Fig. 1 and text for details), as computed by the new algorithm. The black points represent the 400 inversion stations selected.

The model obtained with the Layers procedure is shown in Figure 7, where the same horizontal E-W and N-S vertical cross sections displayed in Figure 6 are shown. There are many similarities to the Growth model regarding the density bodies distributed in the horizontal sections. This interpretation confirms, as shown by BERRINO *et al*. (1998), that the carbonate basement under the volcano (red body in the sections) appears very flat at a depth of about 2.5 km. An uprising of the superimposed layer from about 2 km b.s.l. up to sea level, with a density of 2,400 kg/m^3 to 2,450 kg/m^3 (likely a volcano-sedimentary filling of the Campanian Plain) (BERRINO *et al*., 1998), is clearly detectable in the *b*, *c* and N-S vertical sections, along with a negative density body inside the volcano. Both of these bodies coincide in position with the high and low density bodies already detected through the Growth process, suggesting that they might represent the same structures. Moreover, in the N-S section, there is also an uprising of the carbonate basement. It is surprising that this basement uprising shows the same shape as the high V_P and V_P/V_S anomaly detected under the crater by DE NATALE *et al*. (2004). Also, we should note here that the model was created in an automatic and non-subjective way, and the only initial constraint was the choice of density contrasts.

Finally, to argue about deep sill indicated by AUGER *et al*. (2001), let us analyze a wider area, consisting of 6,203 gravity values (the larger black rectangle in Fig. 1), through both the Growth and Layers processes. The first results here are shown in Figure 8.

Using the Growth process (Fig. 8, left), a large low density body is detectable at a depth of 8 km, which extends to about 12 km, NW of the Vesuvius area. This density body also appears using the Layers approach (Fig. 8, right), although here it appears as a

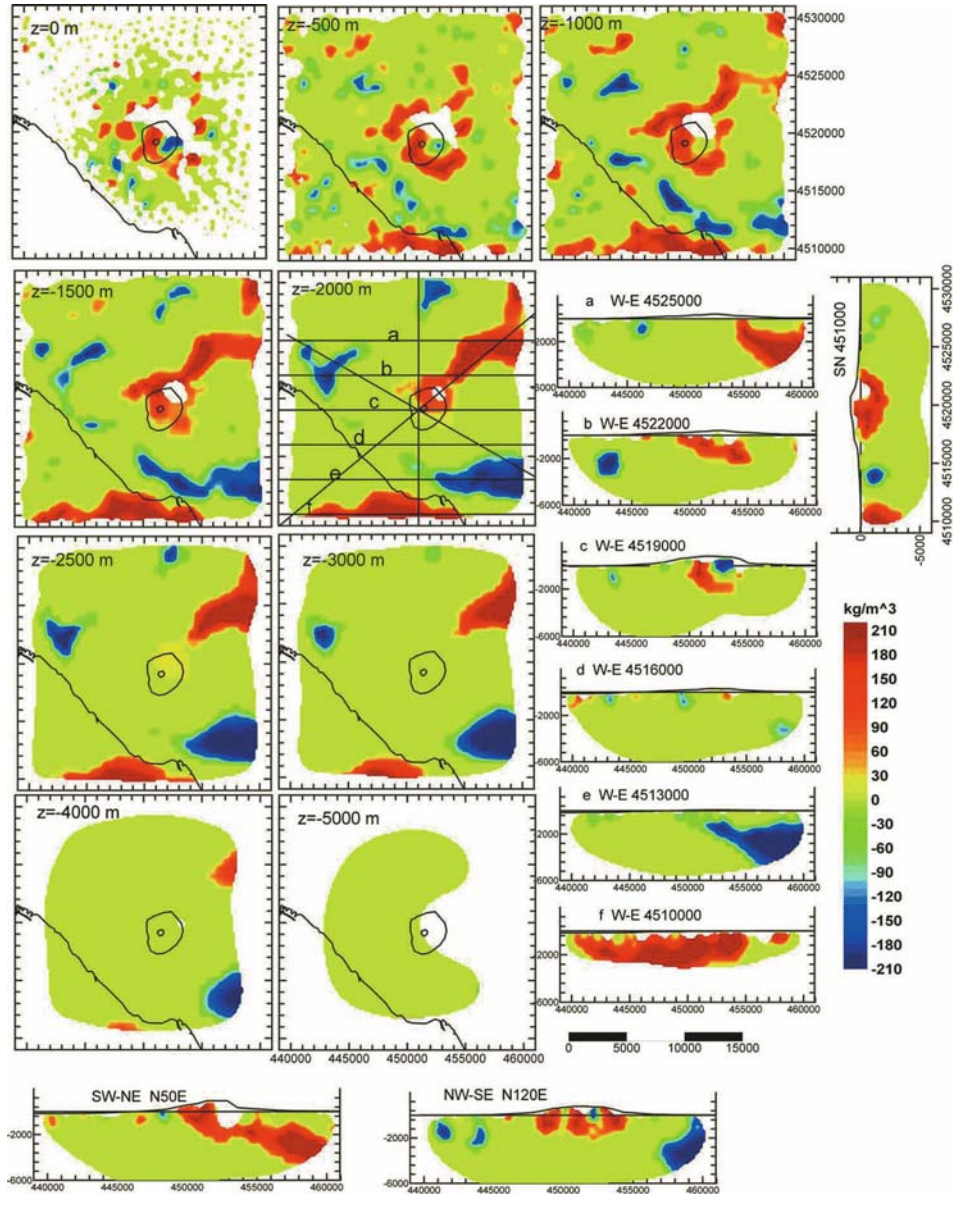


Figure 6
Deep horizontal and vertical *versus* depth cross sections of the inversion model obtained with the cell-Growth process.

body that extends from 8 km down to 10 km, which is elongated mainly in the NW-SE direction and which is in the form of a depression of the top of the crystalline basement (Fig. 8, right, red layers). This depression is evident in the NW-SE vertical profile, but

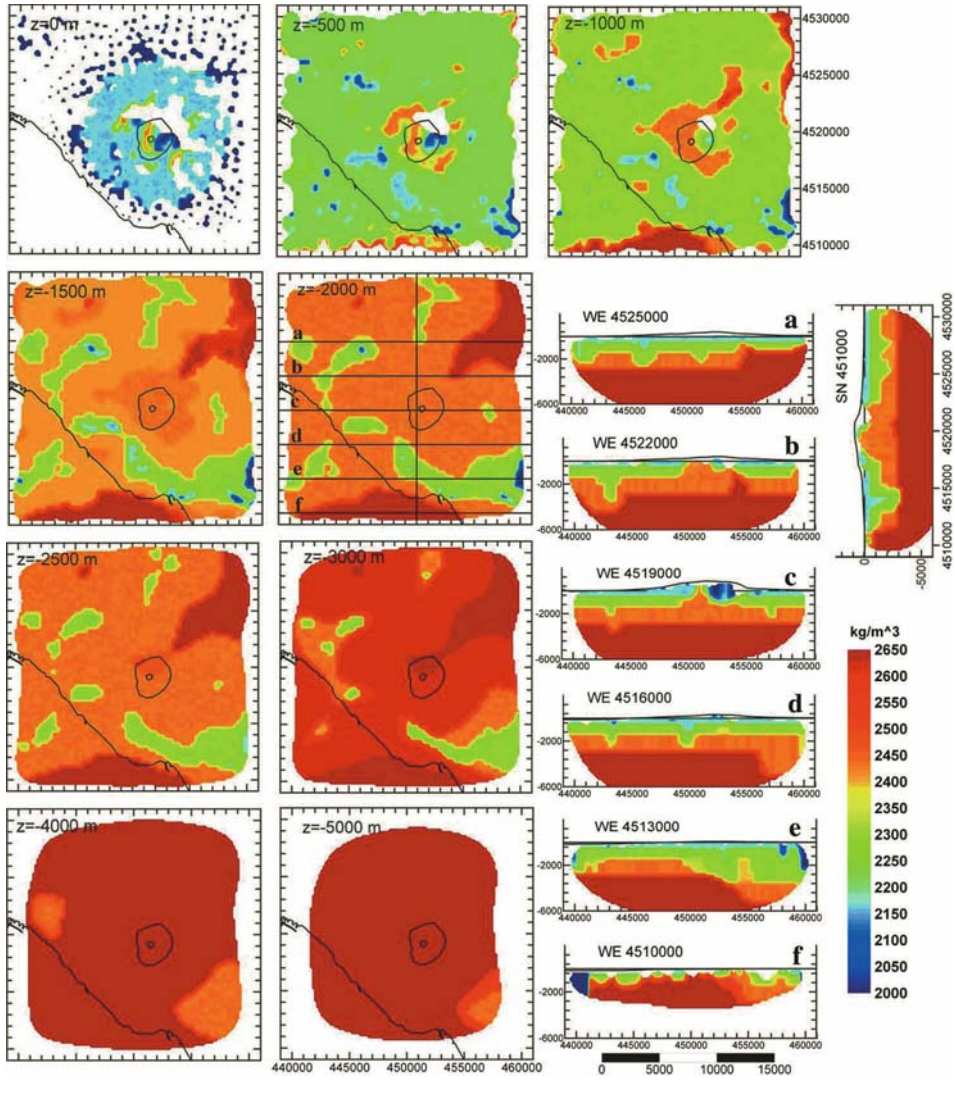


Figure 7
Deep horizontal and vertical *versus* depth cross sections of the inversion model obtained with the shaped-Layers process.

does not appear in the coinciding NE-SW profile analyzed by AUGER *et al*. (2001). The depression detected also corresponds to displacements of the superimposed carbonatic basement (see Fig. 8, right, NW-SE profile, orange layers) that occur in coincidence with the Acerra depression (NW) and the Pompei graben (SE). Based on our current results, the low density body is unlikely to represent the sill-like structure suggested by AUGER *et al*. (2001) and NUNZIATA *et al*. (2006). Our results are more in support of the C shaped negative velocity/density anomaly detected by TONDI and DE FRANCO (2006).

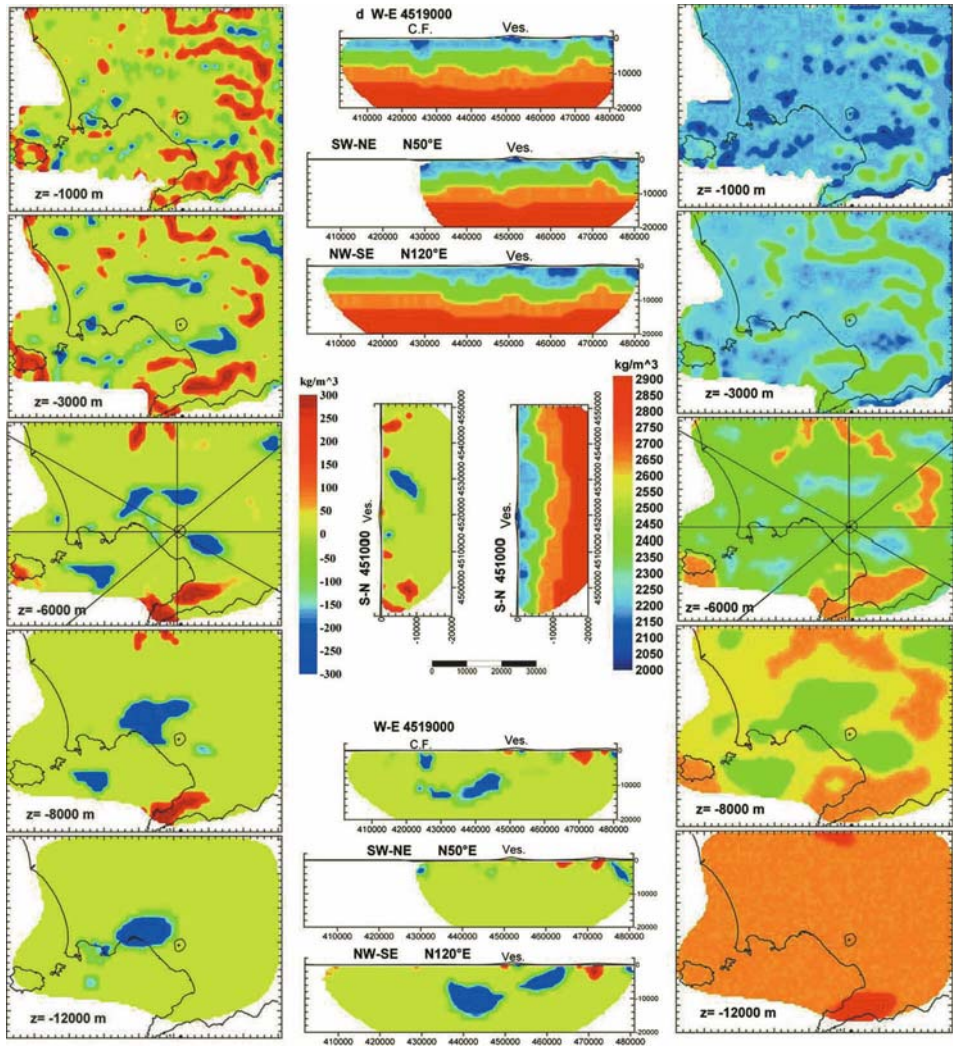


Figure 8
Comparisons along the selected deep horizontal layers (depth from 1 km to 12 km) and the selected vertical *versus* depth profiles crossing the Vesuvius crater, modelled for the whole Neapolitan area with both the Growth and the Layers processes (see text for details).

5. *Conclusions*

We have carried out a 3D interpretation of the available Bouguer gravity anomalies in the Neapolitan area through a new and original algorithm, known as Layers, which was realized to satisfy the aim of our study. This algorithm works in an automatic and non-subjective way, and it has allowed us to define the structural setting below Mt. Vesuvius in a very objective manner, in terms of several layers, each of which represents a specific

geological formation. The same data have also been interpreted in terms of isolated and shallow anomaly density bodies using a well tested algorithm, known as Growth, which also furnished the basic idea for the Layers procedure.

The final models generally confirm the global setting of the area as outlined in previous studies, mainly regarding the shape and the depth of the carbonate basement below Mt.Vesuvius; they also show lateral density contrasts inside the volcano edifice that were only hypothesized in the 2.5D inversion. Moreover, these models have allowed us to indicate a high density body that rises from the top of the carbonate basement and elongates further at sea level, which is probably a rising of the same basement, just below the volcano. As already indicated, the space coincidence of this rising density anomaly with the V_P and V_P/V_S anomaly detected under the crater is surprising. However, since the conversion of seismic velocities to densities and seismic depths is subject to errors, we have to assume some ambiguity in this comparison between the gravity and seismic models.

The results obtained also reveal that the two inversion methods result in very similar models, as the high density isolated body in the Growth model can be associated with the rising high density anomaly in the Layers model. This is supported by comparing the two models through the most significant, in our opinion, selected horizontal (2-km depth) and vertical (SN and WE 4519000 [c] — crossing the Vesuvius crater) profiles (see Figs. 6 and 7).

Taking into account that the density of these modelled bodies, at about 2,400 kg/m^3 to 2,450 kg/m^3, is similar to that assigned to the Vesuvian lavas (2,480 kg/m^3) (Cassano and La Torre, 1987), we suggest that these modelled bodies represent solidified magma bodies, as already indicated by Berrino *et al*. (1998) and De Natale *et al*. (2004).

Finally, with regard to the analysis extended to the entire Neapolitan area to survey the deeper structures, we did not detect any deep bodies that are clearly associable with the sill suggested by Auger *et al*. (2001).

The different ways in which these two algorithms operate (noting that one is aimed at the detection of isolated and shallow masses, the other at the detection of deep subsoil layered structures) is one of the limits of a unique and unambiguous interpretation, in terms of the structural setting provided by the resulting models. This suggests that the fusion of the two algorithms into one will allow the simultaneous modelling of isolated, shallow, deep and layered structures, and this will provide both more information about deep density stratification and a more global vision of the geological and structural setting of the area investigated. We therefore hope to obtain a quasi-univocal model that will be supported by the non-subjective interpretation, and this will be our next tool for further analysis.

Acknowledgments

The authors are very grateful to Professor Gennaro Corrado for stimulating discussions, comments and encouragement. We also wish to give special thanks to G. Jentzsch

and J.H. Gottsmann for their useful comments and suggestions, which have improved the manuscript. This study was funded by INGV-DPC 2005-2006 Projects on Volcanology (Project V3, Sub-Project V3_4 – VESUVIO. Task 1: The volcanic structure and the magma feeding system. UR V3_4/02, Scientific Responsible: G. Berrino), and was developed in part during three stages in Madrid of G.B. This research was also supported by Project GEOMOD (CGL2005-05500-C02).

REFERENCES

ACHAUER, U., BERRINO, G., CAPUANO, P., and DE GORI, P. (2000), *TOMOVES Working Group. Joint inversion of Bouguer gravity and teleseismic delay time data for the Mt. Vesuvius/Campania region's deeper structure: Idea-strategy-first results*, XXV General Assembly of EGS, Nice (France), April 25–29.

ACHAUER, U., BERRINO, G., CAPUANO, P., DE NATALE, G., DESCHAMPS, A., CHIARABBA, C., and GASPARINI, P. (1999), J*oint interpretation of gravity and seismic tomography data for Mt. Vesuvius*, XXII General Assembly of IUGG, Birmingham (UK), July 18–30.

AL-CHALABI, M. (1971), *Some studies relating to non-uniqueness in gravity and magnetic inverse problem*, Geophysics *36* (5), 835–855.

AUGER, E., GASPARINI, P., VIRIEUX, J., and ZOLLO, A. (2001), *Seismic evidence of an extended magmatic sill under Mt. Vesuvius*, Science, *294*, 1510–1512.

BALDUCCI, S., VASELLINI, M., and VERDIANI, G. (1985). *Exploration well in the "Ottaviano" Permit, Italy*. In Strub, A.S., Ungemach, P., Eds.), *European Geothermal Update*, Proc. 3rd Int. Seminar on the Results of EC Geothermal Energy Research, Reidel.

BARBERI, F., INNOCENTI, F., LIRER, L., MUNNO, R., PESCATORE, T., and SANTACROCE, R. (1978), *The Campanian Ignimbrite: A major prehistoric eruption in the Neapolitan area (Italy)*, Bull. Volcanol. *41* (1), 1–22.

BARBOSA, V.C.F., SILVA, J.B.C., and MEDEIROS, W.E. (1997), *Gravity inversion of basements relief using approximate equality constraints on depths*, Geophysics *62* (6), 1745–1757.

BERRINO, G. (1995), *Absolute gravimetry and gradiometry on active volcanoes of Southern Italy*, Bollettino di Geofisica Teorica ed Applicata XXXVII (146), 131–144.

BERRINO, G., CORRADO, G., and LUONGO, G. (1991), *Indagini gravimetriche a mare nelle aree vulcaniche napoletane*. In Proc. 14th Annual Meeting of the Gruppo Nazionale di Geofisica della Terra Solida, pp. 763–775.

BERRINO, G., CORRADO, G., and RICCARDI ,U. (1998), *Sea gravity data in the Gulf of Naples: A contribution to delineating the structural pattern of the Vesuvian area*, J. Volc. Geotherm. Res. *82*, 139–150.

BERRINO, G., CORRADO, G., RICCARDI, U. (2008), *Sea gravity data in the Gulf of Naples: A contribution to delineating the structural pattern of the Campi Flegrei – Ischia sector*, J. Volcanol. Geotherm. Res., doi: 10.1016/j.jvolgeores.2008.03.007.

BERRINO, G., MARSON, I., ORLANDO, L., BALESTRI, L., BALIA, R., BONCI, L., BOZZO, E., CARROZZO, M.T., CERUTTI, G., CESI, C., CIMINALE, M., CRESPI, M., DE MARIA, P., FERRI, F., LODDO, M., LUZIO, D., PINNA, E., and ROSSI, A. (1995), *Rete gravimetrica Italiana di ordine zero. Stato di avanzamento*. In Proc. of the 14th Annual Meeting of the Gruppo Nazionale di Geofisica della Terra Solida, pp. 453–460.

CAMACHO, A.G., MONTESINOS, F.G., and VIEIRA, R. (2000), *A 3-D gravity inversion by means of a growing bodies*, Geophysics, *65*(1), 95–101.

CAMACHO, A.G., MONTESINOS, F.G., and VIEIRA, R. (2002), *A 3-D gravity inversion tool based on exploration of model possibilities*, Comp. Geosci. *28*, 191–204.

CAMELI, G.C., RENDINA, M., PUXEDDU, M., ROSSI, A., SQUARCI, P., and TAFFI, L. (1975), *Geothermal Research in Western Campania (Southern Italy). Geological and geophysical results*, Proc. 2 U.N. Symp. of Development and Use of Geothermal Resources, San Francisco, CA, pp. 315–328.

CARRARA, E., IACOBUCCI, F., PINNA, E., and RAPOLLA, A. (1973), *Gravity and magnetic survey of the Campanian volcanic area, Southern Italy*, Boll. Geof. Teor. ed Appl. *15*, 57, 39–51.

CARROZZO, M. T., LUZIO, D., MARGOTTA, C., and QUARTA, T. (1986), *Gravity map of Italy,* CNR-Progetto Finalizzato Geodinamica.

CASSANO, E. and LA TORRE, P. (1987), *Geophysics.* In (SANTACROCE R., ed.) *Somma Vesuvius*, CNR, Quad. Ric. Sci. *114* (8), 175–196.

CHAKRABORTY, K. and ARGAWAL, B.N.P. (1992), *Mapping of crustal discontinuities by wavelength filtering of the gravity field*, Geophys. Prospect. *40*, 801–822.

CIVETTA, L., D'ANTONIO, M., DE LORENZO, S., DI RENZO, V., and GASPARINI, P. (2004), *Thermal and geochemical constraints on the 'deep' magmatic structure of Mt. Vesuvius*, J. Volcanol. Geoth. Res., *133*, 1–12.

CORTINI, M.R. and SCANDONE, P. (1982), *The feeding system of Vesuvius between 1754 and 1944*, J. Volcanol. Geotherm. Res. *12*, 393–400.

CUBELLIS, E., FERRI, M., and LUONGO, G. (1995), *Internal structures of Campi Flegrei caldera by gravimetric data*, J. Volcanol. Geotherm. Res. *65*, 147–156.

DE NATALE, G., TROISE, C., TRIGILA, R., DOLFI, D., and CHIARABBA, C. (2004), *Seismicity and 3-D substructure at Somma-Vesuvius volcano: Evidence for magma quenching*, Earth and Planet. Sci. Lett., *221*, 181–196.

FERRI, M., CUDELLIS, E., LUONGO, G. (1990), *Strutture crostali del graben della Piana Campana da indagini gravimetriche*, Proc. XI Meeting G.N.G.T.S., C.N.R., I, 737–747.

FEDI, M. (1988), *Spectral expansion inversion of gravity data for 2½D stuctures*, Boll. Geof. Teor. Appl. *21* (121), 25–39.

FINETTI, I. and DEL BEN, A. (1986) *Geophysical study of the tyrrhenian opening*. Boll. Geof. Teor. Appl. XXVIII (11), 75–155.

FINETTI I. and MORELLI C. (1974) *Esplorazione di sismica a riflessione dei Golfi di Napoli e Pozzuoli*. Boll. di Geof. Teor. Appl. *16*, 62–63.

GALLARDO-DELGADO, L.A., PEREZ-FLORES, M.A., and GOMEZ-TREVINO, E. (2003), *A versatile algorithm for joint inversion of gravity and magnetic data*, Geophysics *68*, 949–959.

GÖTZE, H.-J. and LAHMEYER, B. (1988), *Application of three-dimensional interactive modeling in gravity and magnetics*, Geophysics Vol. 53, No. 8, 1096–1108.

LUONGO, G., FERRI, M., CUBELLIS, E., GRIMALDI, M., and ORBIZZO, F. (1988), *Struttura superficiale della Piana Campana: Interpretazione del profilo Garigliano-Campi Flegrei.* Atti del VII Convegno Nazionale G.N.G.T.S., C.N.R., 1121–1128.

MARZOCCHI, W., SCANDONE, R., and MULARGIA, F. (1993), *The tectonic setting of Mount Vesuvius and the correlation between its eruptions and earthquakes of the Southern Appennines*, J. Volcanol. Geotherm. Res. *58*, 27–41.

MORITZ, H. (1984), *Geodetic reference system 1980*, In (Tscherning, ed.), *The Geodesist ' s Handbook*. C.C. Bull. Geod. *58*, 388–398.

NUNZIATA, C., NATALE, M., LUONGO, G., and PANZA, G.F. (2006), *Magma reservoir at Mt. Vesuvius: Size of the hot, partially molten, crustal material detected depper than 8 km*, Earth and Planet. Sci. Lett., *242*, 51–57.

OLIVERI DEL CASTILLO, A. (1966). *Some gravimetric considerations on the Campanian eruptive and sedimentary basin (residual anomalies of (n-1)th order.* Annali Oss. Ves., S *VI*, VIII, 112–137.

PEDERSEN, L.B. (1979), *Constrained inversion of potential field data*, Geophys. Prospect. *27*, 726–748.

PICK, M., PICHA, J., and VYSKÔCIL, V., *Theory of the Earth́s gravity field* (Elsevier, New York (1973)) 538 pp.

RADHAKRISHNA MURTHY, I.V. and JAGANNADHA RAO, S. (1989), *A FORTRAN 77 program for inverting gravity anomalies of two-dimensional basement structures*, Comp. Geosci., *15*-7, 1149–1156.

RAMA RAO, P., SWAMY, K.V., and RADHAKRISHNA MURTHY, I.V. (1999), *Inversion of gravity anomalies of three-dimensional density interface*, Comp. Geosci. *25*, 887–896.

RENÉ, R.M. (1986), *Gravity inversion using open, reject, and "shape-of-anomaly" fill criteria*, Geophysics *51*, 4, 988–994.

ROSI, M., SANTACROCE, R., and SHERIDAN, M.F. (1987), V*olcanic hazard.* In *Somma Vesuvius*, (ed. R. Santacroce) Quad. Ric. Sci. *114*, 197–220.

ROTHMAN, D.H. (1985), *Nonlinear inversion, statistical mechanics, and residual statics estimation*, Geophysics *50*(12), 2784–2796.

ROUSSEEUW, P.J. and LEROY, A.M., *Robust Regression and Outlier Detection* (John Wiley & Sons, New York (1987)) 327 pp.

SANTACROCE, R. (1987), *Somma-Vesuvius*, Quaderni de "La ricerca Scientifica" C.N.R. *114*(8), 251.

SCALA, O. (2002), *Assetto strutturale della Piana Campana dall'analisi di dati gravimetrici*, Ph.D Thesis in Geophysics and Volcanology, University of Naples, Tutors: Corrado G. and Berrino G., Naples.

Silva, J.B.C. and Hohmann, G.W. (1983), *Nonlinear magnetic inversion using a random search method*, Geophysics, *46*(12), 1645–1658.

Tarantola, A., *The inverse problem theory: Methods for data fitting and model parameter estimation* (Elsevier, Amsterdam (1988)) 613 pp.

Tondi, R., and De Franco, R. (2003), *Three-dimensional modeling of Mount Vesuvius with sequential integrated inversion*, J. Geophys. Res. *108*(B5), 2256, doi:10.1029/2001JB001578.

Tondi, R., and de Franco, R. (2006), *Accurate assessment of 3D crustal velocity and density parameters: Application to Vesuvius data sets*, Phys. Earth Planet. Interiors, *159*, 183–201.

Won, J. and Bevis, M. (1987), *Computing the gravitational and magnetic anomalies due to a polygon: Algorithms and Fortran subroutines*, Geophysics *52*, (2), 232–238.

Zollo, A., Gasparini, P., Virieux, J., De Natale, G., Biella, G., Boschi, E., Capuano, P., De Franco, R., Dell'Aversana, P., De Matteis, R., Guerra, I., Iannaccone, G., Le Meur, H., Mirabile, L., and Vilardo, G. (1996), *Seismic evidence for a low velocity zone in the upper crust beneath Mt. Vesuvius*, Science *274*(5287), 592–594.

(Received March 6, 2007, revised April 1, 2008, accepted April 3, 2008)
Published Online First: July 11, 2008

Pure appl. geophys. 165 (2008) 1117–1129
0033–4553/08/061117–13
DOI 10.1007/s00024-008-0349-8

Pure and Applied Geophysics

Testing Logselfsimilarity of Soil Particle Size Distribution: Simulation with Minimum Inputs

Carlos García-Gutiérrez, and Miguel Ángel Martín

Abstract—Particle size distribution (PSD) greatly influences other soil physical properties. A detailed textural analysis is time-consuming and expensive. Soil texture is commonly reported in terms of mass percentages of a small number of size fractions (typically, clay, silt and sand). A method to simulate the PSD from such a poor description or even from the poorest description, consisting in the mass percentages of only two soil size fractions, would be extremly useful for prediction purposes. The goal of this paper is to simulate soil PSDs from the minimum number of inputs, i.e., two and three textural fraction contents, by using a logselfsimilar model and an iterated function system constructed with these data. High quality data on 171 soils are used. Additionally, the characterization of soil texture by entropy-based parameters provided by the model is tested. Results indicate that the logselfsimilar model may be a useful tool to simulate PSD for the construction of pedotransfer functions related to other soil properties when textural information is limited to moderate textural data.

Key words: Soil, particle size distribution, fractals, fragmentation, logselfsimilarity, iterated function system.

1. Introduction

Soil PSD is a fundamental soil property that greatly influences soil porosity and mechanical and hydraulic properties. Its description is usually made for soil particles with sizes smaller than 2 mm. A comprehensive description of PSD within this small size interval requires a sophisticated texture analysis, including novel techniques like laser diffraction analysis (see Montero and Martin, 2003). These analyses have to be repeated for every soil sample, and are highly time-consuming, and expensive.

Attempts to find an equation to simulate the PSD were published in Hatch and Choate (1929), Krumbein and Pettijohn (1938), Otto (1939), Inman (1952). In Buchan *et al*. (1993) several equations for the distribution are compared and the authors showed that the best one was a lognormal model.

Dpto. de Matemática Aplicada, E.T.S.I. Agrónomos, Universidad Politécnica de Madrid, 28040 Madrid, Spain. E-mail: carlos.garciagutierrez@upm.es

Another equation for soil PSD has been derived based on the observed scaling behavior of the number $N(R)$ (or mass) of particles of size greater than a given R: Turcotte, in 1986, showed the scaling rule

$$N(R) \approx R^{-D}, \tag{1.1}$$

D being a number called the *scaling fractal dimension*, which is at present known as *Turcotte's Law*.

Since then considerable work has been devoted to testing the fractality of the soil PSD (see ANDERSON *et al*. 1998, for a review, and TURCOTTE, 1992; TYLER and WHEATCRAFT, 1989; TYLER and WHEATCRAFT, 1992; WU *et al*., 1993 for specific results).

Soil PSD is usually reported by providing only mass percentages of clay (particles with sizes $\leq$ 0.002 mm), silt (0.002-0.05 mm) and sand (0.05-2 mm). A method to simulate the distribution from this poor description would be extremely valuable for further use of PSD for various prediction purposes. To attain that objective, a hypothesis on the distribution is needed.

The power scaling (1.1) implies the fractal behavior of the particle size distribution. This fractal behavior of the PSD inspired the use of mathematical self-similar mass distributions to simulate the entire distribution (MARTIN and TAGUAS, 1998). The self-similarity hypothesis was further tested in TAGUAS *et al*. (1999).

In MARTIN and GARCIA-GUTIERREZ (2006), the model was revised by changing the self-similar hypothesis into a logselfsimilarity assumption about the distribution, based on the fact that the mass of the clay, silt and sand textural fractions is comparable but the size ranges of these fractions (0.002, 0.048 and 1.95 mm) are only comparable in the log-scale. Random logselfsimilar cascades were used to simulate soil distributions beyond the available data, obtaining a surprising result: The best simulation results are attained when the variance of the random factor is close to 0, this is, using only exact logselfsimilarity.

The purpose of this paper is to test the strict logselfsimilarity of particle size distributions by means of iterated function systems: We test the capability to simulate the distributions with the minimum number of textural data inputs, that is, three or even only two soil textural fractions. We make a detailed study using a large dataset (USDA-SCS, 1975) and compare the results with those obtained following the previous self-similar hypothesis (MARTIN and TAGUAS, 1998). We also test the characterization of textures via heterogeneity parameters provided by the logselfsimilar model.

The paper is organized as follows: In section 2 the logselfsimilar model is explained in detail. In section 3, we present the materials which are used to test the model and the way in which the model is applied. Section 4 contains the results and section 5 provides the conclusions of this work.

2. Theory

Soil PSD is viewed as a distribution or measure that assigns to any interval $I = [a, b]$ of $\mathbb{R}$, the mass of soil particles whose size (equivalent diameters) is greater or equal to a and less or equal to b. Next we present theory related with self-similar mass distributions and the logselfsimilar model.

2.1. Selfsimilar Mass Distributions (Measures)

Given a set of functions (linear transformations)

$$\varphi_i : \mathbb{R} \to \mathbb{R}, \quad |\varphi_i(x) - \varphi_i(y)| = r_i|x - y|, \quad r_i < 1, \quad i = 1 \ldots m$$

and a set of positive numbers (probabilities) $p_i \geq 0$, $i = 1 \ldots m$, $\sum_{i=1}^{m} p_i = 1$, a unique mass distribution μ exists such that $\mu(\mathbb{R}) = 1$ and

$$\mu(J) = \sum p_i \mu(\varphi_i^{-1}(J)) \tag{2.1}$$

for $J \subset \mathbb{R}$ (Hutchinson, 1981).
The set $\{ \varphi_i, p_i, i = 1 \ldots m \}$ is called the *iterated function system* (IFS) and the mass distribution is said to be the *selfsimilar mass distribution* of the IFS. The support of the above distribution is the set I which verifies that

$$I = \bigcup_{i=1}^{m} \varphi_i(I). \tag{2.2}$$

Moreover, one has that

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} f(x_k) = \int_{\mathbb{R}} f(x) d\mu(x) \tag{2.3}$$

for all continuous functions $f : \mathbb{R} \to \mathbb{R}$, being $x_k = \varphi_k \circ \cdots \varphi_2 \circ \varphi_1(x_0)$, for all $x_0 \in \mathbb{R}$. In particular this implies that, if $I \subset \mathbb{R}$ is an interval and $m(n)$ is the number of points of $\{x_0, \ldots, x_k\} \cap I$, then (see Martin and Taguas, 1998)

$$\mu(I) = \lim_{n\to\infty} \frac{m(n)}{n+1}. \tag{2.4}$$

Self-similar measures are common examples of so-called *multifractal measures*, that is, measures for which the local *Hölder exponent* of μ at x defined by the limit

$$\alpha(x) = \lim_{r\to 0} \frac{\log \mu(I_r(x))}{\log r} \tag{2.5}$$

is not constant on the support, $I_r(x)$ being the real interval $[x - r, x + r]$.

The Hölder exponents for self-similar measures typically span the entire interval between two extreme values $\alpha_{\min}$ and $\alpha_{\max}$ (see Everstz and Mandelbrot, 1992 or

FALCONER, 1997, for further details). The exponent provides a measure of mass concentration around the point: The greater $\alpha(x)$ is, the smaller will be the mass concentration and *vice versa*. However, for simulated self-similar measures (or experimental measures with self-similar characteristics), the above theoretical approach is replaced with a coarse version involving a scaling analysis of overall information quantities instead of the pointwise local Hölder exponents lacking practical sense in a natural setting. One common choice is to consider dyadic scaling down (EVERSTZ and MANDELBROT, 1992), that is, succesive partitions of I of size $L{\cdot}\varepsilon = L{\cdot}2^{-k}$, L being the length of I and $k = 1, 2, 3,...$ At every size scale ε, a number $N(\varepsilon) = 2^k$ of subintervals (cells) I_i, $i = 1 \ldots N(\varepsilon)$ are considered and their respective measures $\mu(I_i) = \mu_i(\varepsilon)$ assumed to be provided by available data. Now, the ratio $\log \mu_i(\varepsilon)/\log\varepsilon$ is called the coarse Hölder exponent of interval I_i and the *Hölder spectrum* is defined via a parameter q such that

$$\alpha(q) \approx \frac{\sum_{i=1}^{N(\epsilon)} \mu_i(q,\epsilon) \log \mu_i(\epsilon)}{\log \epsilon}, \tag{2.6}$$

where

$$\mu_i(q,\epsilon) = \frac{\mu_i(\epsilon)^q}{\sum_{i=1}^{N(\epsilon)} \mu_i(\epsilon)^q}$$

and "$\approx$" means that a suitable linear fitting holds for a range of scales (ε values) where we want to characterize the scaling regularity of the measure μ (see EVERSTZ and MANDELBROT, 1992; CHHABRA and JENSEN, 1989 for details).

For exact self-similar measures, "$\approx$" in equation (2.6) may be replaced with the limit when $\epsilon \to 0$. In such a case, the function $\alpha(q)$ for $-\infty < q < +\infty$ parameterizes the interval $[\alpha_{\min}, \alpha_{\max}]$ of local Hölder exponents.

On the other hand, a suitable fitting of equation (2.6) applied to experimental data, for a certain range of scales, may reveal that the measure concerned has self-similar features within that range.

2.2. *The Logselfsimilar Model*

Soil PSD is defined by assigning to each interval $I = [a, b] \subset \mathbb{R}$ the mass $\mu(I)$ of particles whose size is in that interval. This distribution can be seen as the result of a fragmentation process, this is, an iterative processs acting within a range of scales. Experimental data on this distribution showed power scaling of the type reflected in Section 1 (ANDERSON *et al.*, 1998). This fact suggests that the distribution should have scale-invariant behavior: If we zoom into the mass interval to see it at a finer scale, it should resemble (statistically) the structure of the whole interval: A photograph of soil looks similar at every scale; it is impossible to guess the size of elements in the picture.

Once the invariance with respect to the scale becomes a sensible hypothesis, the problem that arises is to determine how this invariance-based model can be used. MARTÍN

and TAGUAS (1998) proposed a self-similar model generated by iterated function systems (IFS) that was useful for simulating self-similar PSDs from the knowledge of common textural data (clay, silt and sand mass proportions). This model generates a self-similar mass distribution via an iterative process that allocates the relative mass proportions of the elementary size classes in reduced linear copies of the size interval. Testing this model showed that the use of the clay (soil particles smaller than 0.002 mm.), silt (2–50 mm) and sand (50–2000 mm) fractions, for example, as inputs for the model lead to a very unrealistic simulated PSD. The reason is that these three fractions (subintervals of the mass size distribution) contain similar amounts of mass, however the respective sizes of the intervals differ by orders of magnitude (viz. 0.002 mm, 0.048 mm and 1.95 mm, respectively). The simulations lead to vast amounts of soil mass accumulated in very small linear copies of the size interval (specifically in the reduced linear copies of the clay interval), which contradicts common pedological knowledge.

MONTERO and MARTIN (2003) computed the Hölder spectrum of soil texture data obtained with laser diffraction, and the scaling behavior was excellent when the interval of sizes was log-rescaled. In fact, using the log-rescaled interval instead of the usual interval in scaling analysis is strongly supported by the nature of the data provided by texture analysis instruments (see MARTÍN *et al*., 2001; MONTERO and MARTÍN, 2003).

These facts suggested a reconsideration of the selfsimilar model (MARTIN and GARCIA-GUTIERREZ, 2006). The key idea is to view the PSD as the result of an iterative process that spreads the mass in the log-rescaled particle size interval. Such property is called logselfsimilarity. In MARTIN and GARCIA-GUTIERREZ (2006) random cascades were used to simulate the PSD. The random factor of the model agglomerates all the causes different to logselfsimilarity that could explain the PSD. When variance is 0, the random factor disappears, and logselfsimilarity is the only explanation for the soil PSD. The best simulation results were attained when the random factor was 0, therefore supporting the logselfsimilar behavior of the distribution. This fact lead us to using strict logselfsimilarity and simulating the PSD by utilizing the iterated function system (IFS).

Below we describe an IFS simulation for a PSD based on the logselfsimilarity hypothesis. It is a simple algorithm by which we can obtain the mass of soil particles with sizes within a given interval $I \subset I_0$, being $I_0 = [0,2000]$ the textural interval.

Let p_i be the mass proportions of soil particles corresponding to the size fractions I_i, $i = 1 \ldots q$. Thus $I_0 = \cup_{i=1}^{q} I_i$.

Let ξ_i be the linear transformations that map I_0 into I_i. That is, if $I_i = [a, b]$, then $\xi_i(x) = x(b - a)/2000 + a$. Also let $I_i^* = \Phi(I_i) = [\log (1 + a), \log (1 + b)]$. The new linear transformations, φ_i, are the ones that transform I_0^* into I_i^*, with the same probabilities p_i, $i = 1, 2, 3$.

Then begins the iteration procedure:

(1) Take any starting point x_0 from the support I_0^*.
(2) Choose randomly, with probability p_i, one of the three linear transformations φ_i, $i = 1, 2, 3$ and calculate the next point of the simulation: $x_1 = \varphi_i(x_0)$.

(3) Continue the process as in (2), obtaining all the points of the simulation: $x_k = \varphi_i(x_{k-1})$, with probability p_i, chosen randomly, $i = 1, 2, 3$.

This process defines a limit measure that is multifractal.

With the points $x_0, x_1, \ldots, x_n$ we can obtain the measure at any interval $I \subset I_0$, $\mu(I)$, by

$$\mu(I) = \lim_{n\to\infty} \frac{m(n)}{n+1}$$

$m(n)$ being the number of points of the orbit x_i that fall within the interval I.

For this we have to calculate $I^* = \Phi(I)$ and count $m^*(n)$, the number of points of the orbit that fall within $I^* = \Phi(I) \subset I_0^*$. Then we calculate $\mu\,(I^*) = \mu\,(I)$.

The estimate of $\mu(I)$ is obtained very quickly in the practice, since the convergence of the algorithm is extremely rapid. In fact, computed $\mu(I)$ did not change after $n = 3000$.

2.3. Heterogeneity Parameters

Soil PSD is used in most pedotransfer functions, that is, functions that estimate certain soil properties that are difficult or expensive to measure, in an indirect way via empiric correlations (Van Genuchten and Leij, 1992; Wösten *et al*., 2001). For example, soil hydraulic properties are estimated by using parameters that characterize the shape of the PSD. Therefore PSD characterization and subsequently soil textral classification is an important issue in soil sciences.

The USDA textural triangle is the most common way to classify soil textures. It uses the standard PSD available data (clay, silt and sand mass fractions) to classify the soils in 13 different types (textural classes), according to specific mass fraction boundaries for each class. Other classification systems (Folk, 1954; Shepard, 1954; Baver *et al*., 1972; Vanoni, 1980) follow the above scheme, only with variations of the mass fraction boundaries, however these classifications are rather poor because soils with very different physical properties may fall under the same class.

Also, the use of previously mentioned fractal dimensions in soil classification proved useless, as texturally different soils can have the same dimension. There is a need to develop additional parameters to characterize soil structure that might be better predictors of soil properties.

The entropy dimension is a parameter that measures and characterizes the degree of heterogeneity of a complex distribution. Is it difficult to obtain this parameter directly from a distribution, but a well-known result from fractal geometry (Young, 1982; Deliu *et al*., 1991) allows one to compute this parameter with a simple formula when the distribution is (or is assumed to be) self-similar. This assumption, earlier used to describe and simulate the PSD, was later used in Martin *et al*. (2001) to parameterize the soil texture with entropy dimensions. The textural triangle regions were changed for intervals of values of entropy dimensions, thus obtaining a continuous parameterization of soil texture.

MARTIN *et al*. (2005) used the balanced entropy, which corrects the distortion of the entropy when the size of the intervals are not equal, to parameterize soil texture and predict soil volumetric water content.

Hölder exponents provide information about the mass of particles with sizes within various ranges, and can be related to physical or hydraulic properties related to the packing of particles, like soil water retention. These heterogeneity parameters can be obtained from available texture data, using the model's logselfsimilarity hypothesis.

The Hölder spectrum $\alpha(q)$ of the measure v on the rescaled interval is given by (see FALCONER, 1994, 1997)

$$\alpha(q) = \frac{\sum_{i=1}^{m} p_i^q r_i^\beta \log p_i}{\sum_{i=1}^{m} p_i^q r_i^\beta \log r_i},$$

where $\beta = \beta(q)$ is a positive number verifying

$$\sum_{i=1}^{m} p_i^q r_i^{\beta(q)} = 1.$$

The value

$$\alpha(0) = \frac{\sum_i r_i^* \log p_i}{\sum_i p_i \log r_i^*}$$

would approach the average value of the coarse Hölder exponents for fine partitions of the size interval and

$$\alpha(1) = \frac{\sum_i p_i \log p_i}{\sum_i p_i \log r_i^*}$$

is consistent with the *entropy dimension* of the distribution, v, mentioned above.

These heterogeneity parameters, obtained through the new logselfsimilarity hypothesis (MARTIN and GARCIA-GUTIERREZ, 2006) can be used to characterize the PSD heterogeneity, to quantitatively classify the soils, and potentially to estimate soil hydraulic properties.

3. Materials and Methods

The data used to simulate and test the logselfsimilar theory corresponded to the upper two horizons of soils reported by the Soil Conservation Service (1975). Soil data included the mass proportions m_i of particles in eight size classes (mm): clay (< 0.002), silt (0.002–0.02) and (0.02–0.05), very fine sand (0.05–0.1), fine sand (0.1–0.25), medium sand (0.25–0.5), coarse sand (0.5–1) and very coarse sand (1–2). In order to use these data to construct an IFS, we shall denote by $[a, b]$ the particles with sizes greater than or equal

to a and less or equal to b. These size classes determine a set of seven intermediate cutoff points 0.002, 0.02, 0.05, 0.1, 0.25, 0.5, 1 and eight consecutive intervals corresponding to the eight size classes $I_1 = [0, 0.002]$, $I_2 = [0.002, 0.02]$, ..., $I_8 = [1, 2]$.

These data offer the possibility of using some of them as inputs for the model and simulate a fractal soil PSD associated with it. The simulated data then can be compared with the real data, not used as input for the model to estimate the goodness-of-fit of the simulation.

We also try to find which intervals, from all available in one soil, used as input values for the model, yield the best simulation results.

The number of linear transformations in the simulation vary from 2 to 8, according to the available data. With 2 linear transformations the number of cutoff points is 7, thus there are 7 different input value possibilities. With 3 linear transformations the number of input value possibilities is 21; with 4 linear transformations it is 35. Thus, the method provides a great number of potential simulated PSDs.

To run most of the tests on the model we used 3 linear transformations. The reason for this choice is that three is the number of the most commonly available textural data, namely: the mass percentages of clay, silt and sand. In some cases additional simulations were made following TAGUAS *et al*. (1999), in order to compare those results to the ones obtained with the logselfsimilar cascade model. Both simulation methods use the same soil data as inputs and for comparison with their simulated counterparts.

The model was also tested with two linear transformations and the results were compared with the use of three linear transformations.

We constructed the three linear transformations with different possible logselfsimilar IFS

$$\{\varphi_1, \varphi_2, \varphi_3; p_1, p_2, p_3\}$$

by using the following procedure:

(1) Select two cutoff points α and β among the seven possible choices, $\alpha < \beta$.
(2) Let p_1, p_2 and p_3 be the mass proportions of the three constructed intervals $I_1 = [0, \alpha]$, $I_2 = [\alpha, \beta]$ and $I_3 = [\beta, 2000]$.
(3) Obtain the log-rescaled intervals $I_i^* = \Phi(I_i)$, and assign them the same probability as to the initial intervals. Also calculate the set $I^* = \Phi(I)$
(4) Let φ_i be the linear transformation which maps the interval I^* into I_i^*.

These rules permit us to make up to 21 IFS simulations for each soil, depending on the values of the two cutoff points (α and β). Another 21 simulations were performed in accordance with the self-similar scheme of TAGUAS *et al*. (1999). The results obtained with these methods were compared via the error of the respective simulations, defined as

$$\epsilon = \frac{\sum |m_i - m_i'|}{2}, \tag{3.1}$$

m_i being the real mass proportion in the size class I_i, and $m_i^{'}$ the mass proportion assigned to the same size class by the simulation. This error was used in TAGUAS *et al*. (1999), where the self-similar model was tested, therefore we use the same formula in order to keep coherence on the testing method. A similar formula was employed earlier in VRSCAY (1991). The rationale for dividing by two 3.1 in is the following: the formula adds mass deviations for the intervals. A positive mass deviation in one interval derives from a negative one in another, and *vice versa*. In some way, mass deviations are accounted two times, therefore we divide by two.

For each soil 21 different simulations were created with the different input value possibilities, and 21 different error values were obtained. We used the minimum one as the error of the PSD simulation because it corresponds to the best simulation result with the logselfsimilar model (using only three data).

The above scheme varies slightly when using 2 linear transformations (instead of 3). In this case there is only one possible cutoff point choice, and the number of intervals is only 2. The number of possible IFSs per soil is 7 when the first interval starts at zero, but we also tested the IFS when the first interval starts at 0.002, this is ignoring the clay fraction. The number of possible IFs in this case is 6, which is the number of possible cutoff points {0.05, 0.1, 0.25, 0.5, 1}.

The error in this case was also calculated with 3.1.

4. Results and Discussion

Textural data of 171 soils have been studied. We excluded soils whose mass proportions were polarized into any of the three classes (silt, clay or sand). Therefore, the soils whose clay and sand content was more than 85% and those whose silt content was more than 90% were eliminated from the list. The number of remaining soils was 158.

First we used three linear transformations (3 subintervals or mass proportions) as input values for the model to compare it to the previous self-similar model (TAGUAS *et al*., 1999). For 111 soils (70.3% of the selected soils) the error was smaller in the logselfsimilar model than in the self-similar.

The average error of the logselfsimilar simulations was 10.9 whereas the average error for the self-similar ones was 14.7.

With 3 linear transformations the error for each soil is the minimum of the errors of the simulations with the 21 different input value possibilities. For the self-similar case the minimum is attained with the 0-0.002-0.020-2 partition as input values on 90 out of the 158 soils (57%). The next partition with the most minimums is 0-0.05-0.1-2. For the logselfsimilar case the minimum is attained with the partition 0-0.002-0.02-2 in 34 soils (21.5%) and with the partition 0-0.002-0.5-2 in the other 22 soils (13.9%).

Another test was to measure the error on all the soils for the clay, silt and sand fractions, which are the most readily available soil data, in the logselfsimilar model

and in the self-similar model. Clay particles have diameters between 0 and 0.002 mm, silt particles have diameters between 0.002 and 0.02 mm in the ISSS classification, and between 0.002 and 0.05 mm in the USDA classification. The particles with sizes between the upper limit of the silt and 2 mm are considered to be sand particles. Therefore the two input values tested were 0-0.002-0.02-2 (ISSS) and 0-0.002-0.05-2 (USDA). In the ISSS partition case the mean error was 20.0 for both cases. In 93 soils (58.9%) the error was smaller in the logselfsimilar simulation. For the USDA partition the mean error was 29.6 for the self-similar case and 23.9 for the logselfsimilar case. In 127 soils (80.3%) the error was smaller in the logselfsimilar simulation of the soil PSD.

By using 2 initial linear transformations (two input values), instead of 3, we compared the results with the self-similar model using the same number of linear transformations. This test was done in two ways: with the first interval starting from size 0 and with the first interval starting from size 0.002, this is, not counting the clay subinterval. The mean value of the error of all the soils tested (158, the selection mentioned above) was 22.3 for the logselfsimilar model and 30.3 for the selfsimilar one, when the first interval contained the clay fraction. When the clay fraction was not included in the simulation, this is, when the first interval started at 0.002 mm., the mean value of the error was 37.7 for the logselfsimilar case and 41.2 for the self-similar case.

The results demonstrate that the logselfsimilarity hypothesis is more realistic than the previous self-similar hypothesis when using only two linear transformations. The results also show that the use of two linear transformations yields greater errors than when using 3 linear transformations. When increasing the number of inputs in the model, soil PSDs are simulated better. The number of input values depends on the textural fractions that are known from the soil, which is usually three. The results also show that the soil particles in the clay fraction are also involved in the fractal logselfsimilar structure of the PSD. This is a surprising and not expected result because clay particles are presumably formed not as a result of a fragmentation process, which is the explanation for the fractal logselfsimilar structure of the PSD.

The Hölder spectrum $\alpha(q)$ for all soils was calculated for several values of q. Table 1 shows the mean values of $\alpha(0)$, $\alpha(1)$ and $\alpha(1)/\alpha(0)$ for the different textural classes of the USDA textural triangle in all 171 soils. For this parameter the Loam, Sandy Loam and Sandy Clay Loam classes are not distinguishable. The same is true for the Silt Loam and Silty Clay Loam classes. Nonetheless, when varying the parameter q the Hölder spectrum is capable of distinguishing between those classes: Table 2 shows the Hölder parameters for the previous classes with $q = 2, -2, 10, -10$.

The heterogeneity parameter wellness depends on the chosen model. The better the model, the more accurately the parameters will quantitatively characterize the texture and, in addition, the better to establish regressions with soil physical properties or build new pedotransfer functions.

Table 1

Mean values of α(0), α(1) and α(1)/α(0) for the different textural classes

Textural Class	α(0)	α(1)	α(1)/α(0)
Sand	1.65954067	0.624458	0.37994797
Loamy Sand	1.3023216	0.7883644	0.61170227
Sandy Loam	1.14006598	0.89193242	0.9493417
Silt Loam	1.37489666	0.77902314	0.59723171
Silt	1.56655933	0.646987	0.4217506
Loam	1.14928982	0.90012724	0.79933527
Sandy Clay Loam	1.137707	0.8940058	0.79315775
Clay Loam	1.19510946	0.87178831	0.73544949
Silty Clay Loam	1.3627452	0.77143413	0.58778276
Silty Clay	1.43659986	0.71459671	0.52387275
Clay	1.42989975	0.60704275	0.4406898

Table 2

Mean values of α(2), α(−2), α(10) and α(−10) for some textural classes

Textural Class	α(2)	α(−2)	α(10)	α(−10)
Sandy Loam	0.7787254	1.68533752	0.648082	1.82205762
Silt Loam	0.67265007	2.22231493	0.56712528	2.33494603
Loam	0.80783118	1.69722282	0.68345041	1.81338135
Sandy Clay Loam	0.785247	1.678321	0.6553278	1.7969594
Silty Clay Loam	0.65487927	2.03840833	0.56835773	2.15139607

5. Conclusions

The fractal logselfsimilar hypothesis for the structure of the soil PSD was tested in this work. The new model was compared to the previous self-similar one on quality data and was found to substantially improve the simulation of the soil PSD.

When using 3 input values the errors of the logselfsimilar simulations were smaller than the errors when using the self-similar model for 70% of soils and the average error was 10.9, in contrast to 14.7, which was the average error with the previous model. The error mean values when using just two input values were 22.3 and 30.6, respectively.

The error of the simulations depended greatly on the input values used. The best input values for the logselfsimilar model are the mass of particles with sizes in the following intervals [0, 0.002], [0.002, 0.02] and [0.02, 2], which correspond to the clay, silt and sand fractions under the ISSS classification.

The heterogeneity parameters provided by the model can quantitatively characterize soil texture and may be used to build new pedotransfer function or be related to soil physical properties related to PSD.

Acknowledgements

This work has partially been supported by Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I) under ref. AGL2007-62648 Spain and by DGUI (Comunidad de Madrid) - UPM (Technical University of Madrid), Ref.:M0700204135. We thank the reviewers for their valuable comments and suggested changes. They significantly helped us to enhance the paper.

References

Anderson, A.N., McBratney, A.B., and Crawford, J.W. (1998), *Applications of fractals to soil studies*, Adv. Agron. *63*, 1–76.

Barnsley, M.F. and Demko, S. (1985), *Iterated function systems and the global construction of fractals*, Proc. R. Soc. Lond. A *399*, 243–275.

Barnsley, M.F., Erwin, V., Hardin, D., and Lancaster, J. (1985), *Solution of an inverse problem for fractals and other sets*, Proc. Natl. Acad. Sci. U.S.A. *83*, 1975–1977.

Baver, L.D., Gardner W.H., and Gardner, W.R. *Soil physics* (4th edn. Wiley, NY 1972).

Buchan, G.D., Grewal, K.S., and Robson, A.B. (1993), *Improved models of particle-size distribution: An illustration of model comparison techniques*, Soil. Sci. Soc. Am. J. *57*, 901–908.

Chhabra, A. and Jensen, R. V. (1989), *Direct determination of the f(α) singularity spectrum*, Phys. Rev. Lett. *62*, 1327–1330.

Deliu, A., Geronimo, J.S., Shonkwiller, R., and Hardin, D. (1991), *Dimensions associated with recurrent self-similar sets*, Math. Proc. Camb. Phil. Soc. *110*, 327–336.

Everstz, C.J.G and Mandelbrot, B.B. *Multifractal measures* In *Chaos and Fractals* (eds. Peitgen, H., Jürgens H., and Saupe D.) (Springer, Berlin, 1992) pp. 921–953.

Falconer, K.J. (1994), *The multifractal spectrum of statistically self-similar measures*, J. Theoret. Prob. *7* (3), 681–702.

Falconer, K.J. *Techniques in Fractal Geometry* (Wiley & Sons, Chichester 1997).

Folk, R.L. (1954), *The distinction between grain size and mineral composition in sedimentary rock nomenclature*, J. Geol. *62*, 344–359.

Hatch, T. and Choate, S. (1929), *Statistical description of the size properties of nonuniform particulate substances*, J. Franklin Inst. *207*, 369–387.

Hutchinson, J. (1981), *Fractals and selfsimilarity*, Indiana Univ. J. of Mathematics *30*, 713–747.

Inman, D.L. (1952), *Measures for describing the size distribution of sediments*, J. Sediment. Petrol. *22* (3), 125–145.

Krumbein, W.C. and Pettijohn, F.J. *Manual of Sedimentary Petrography* (Appleton-Century, New-York 1938).

Mandelbrot, B.B. *The Fractal Geometry of Nature* (New York, W.H. Freeman, 1982).

Martin, M.A. and Taguas, F.J. (1998), *Fractal modelling, characterization and simulation of particle-size distributions in soil*, Proc. R. Soc. Lond. A *454*, 1457–1468.

Martin, M.A., Rey, J.M., and Taguas, F.J. (2001), *An entropy-based parameterization of soil texture via fractal modelling of particle-size distribution*, Proc. R. Soc. Lond. A *457*, 937–947.

Martin M.A., Pachepsky, Y.A., Rey, J.M., Taguas, J., and Rawls, W.J. (2005), *Balanced entropy index to characterize soil texture for soil water retention estimation*, Soil Sci. *170*,(10).

Martin, M.A. and Garcia-Gutierrez, C. (2008), *Log selfsimilarity of continuous soil particle-size distributions estimated using random multiplicative cascades*, Clays and Clay Minerals, *56* (3), 389–395.

Montero, E. and Martin, M.A. (2003), *Hölder spectrum of dry grain volume-size distributions in soil*, Geoderma *112*, 197–204.

Otto, G. (1939), *A modified logarithmic probability graph for interpretation of mechanical analysis of sediments*, J. Sediment. Petrol. *9* (2), 62–76.

Shepard, F.P. (1954), *Nomenclature based on sand-silt-clay ratios*, J. Sediment. Petrol. *24*, 151–158.

SHIRAZI, M.A. and BOERSMA, L. (1984), *A unifying quantitative analysis of soil texture: Improvement of precision end extension of scale*, Soil. Sci. Soc. Am. J. *52*, 181–190.

TAGUAS, F.J., MARTIN, M.A., and PERFECT, E. (1999), *Simulation and testing of self-similar structures for soil particle-size distributions using iterated function systems*, Geoderma. *88*, 191–203.

TURCOTTE, D.L. (1886), *Fractals and Fragmentation*, J. Geophys. Res. *91*, 1921–1926.

TURCOTTE, D.L., *Fractals and Chaos in Geology and Geophysics* (Cambridge University Press, Cambridge 1992).

TYLER, S.W. and WHEATCRAFT, S.W. (1989), *Application of fractal mathematics to soil water retention estimation*, Soil Sci. Soc. Am. J. *53*, 987–996.

TYLER, S.W. and WHEATCRAFT, S.W. (1992), *Fractal scaling of soil particle-size distributions: analysis and limitations*, Soil Sci. Soc. Am. J. *56*, 362–369.

Young (1982)

VAN GENUCHTEN, M.T.H. and LEIJ, F., *On estimating the hydraulic properties of unsaturated soils*. In Proc. Int. Workshop on *Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils* (eds. van Genuchten, M.T.H., Leij, F.J., and Lund, L.J.) (University of California, Riverside, CA 1992).

VANONI, V.A., *Sedimentation Engineering* (Sedimentation Committe, Hydraulic Division, American Society of Civil Engineers, New York 1980).

VRSCAY, E.R., *Moment and collage methods for the inverse problem of fractal construction with iterated function systems*. In *Fractals in the Fundamental and Applied Sciences* (eds. Peitgen, H.O., Henriques, J.M., and Penedo, J.L.) (North Holland 1991) pp. 443–461.

WÖSTEN, J.H.M., PACHEPSKY, J.A., and RAWLS, J.W. (2001), *Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics*, J. Hydrol. *251*, 123–150.

WU, Q., BORKOVEC, M., and STICHER, H. (1993), *On particle size distributions in soils*, Soil Sci. Soc. Am. J. *57*, 883–890.

YOUNG, L.S. (1982), *Dimension, entropy and Liapunov exponents*, Ergod. Th. Dynam. Sys. *2*, 109–124.

(Received April 2, 2007, revised March 25, 2008, accepted March 27, 2008)
Published Online First: July 11, 2008

Pure appl. geophys. 165 (2008) 1131–1151
0033–4553/08/061131–21
DOI 10.1007/s00024-008-0350-8

Pure and Applied Geophysics

Steric Sea-Level Change and its Impact on the Gravity Field caused by Global Climate Change

SABINE ROEDELSPERGER,[1] MICHAEL KUHN,[2] OLEG MAKARYNSKYY,[2,3] and CARL GERSTENECKER[1]

Abstract—It is sometimes assumed that steric sea-level variations do not produce a gravity signal as no net mass change, thus no change of ocean bottom pressure is associated with it. Analyzing the output of two CO_2 emission scenarios over a period of 2000 years in terms of steric sea-level changes, we try to quantify the gravitational effect of steric sea-level variations. The first scenario, computed with version 2.6 of the Earth System Climate Model developed at the University of Victoria, Canada (UVic ESCM), is implemented with a linear CO_2 increase of 1% of the initial concentration of 365 ppm and shows a globally averaged steric effect of 5.2 m after 2000 years. In the second scenario, computed with UVic ESCM version 2.7, the CO_2 concentration increases quasi-exponentially to a level of 3011 ppm and is hold fixed afterwards. The corresponding globally averaged steric effect in the first 2000 years is 2.3 m. We show, due to the (vertical) redistribution of ocean water masses (expansion or contraction), the steric effect results also in a small change in the Earth's gravity field compared to usually larger changes associated with net mass changes. Maximum effects for computation points located on the initial ocean surface can be found in scenario 1, with the effect on gravitational attraction and potential ranging from 0.0 to $-0.7 \cdot 10^{-5}$ m s^{-2} and $-3 \cdot 10^{-3}$ to $6 \cdot 10^{-3}$ m^2 s^{-2}, respectively. As expected, the effect is not zero but negligible for practical applications.

Key words: Global warming, climate model, sea level change, steric effect, gravity field change.

1. Introduction

Climate change and global warming due to increased emissions of carbon dioxide (CO_2) and other Greenhouse gases have a direct impact on global sea level (e.g., SOLOMON *et al*., 2007). Melting ice and snow, changes in the hydrologic cycle and the warming of ocean water alter the global mean sea level (e.g., DOUGLAS and PELTIER, 2002; HOUGHTON, 2004; BINDOFF *et al*., 2007). While the inflow of melt water from land-based ice masses and the exchange with terrestrial water reservoirs directly influence the total ocean water mass, ocean warming/cooling and freshening/salinification, the steric effect, causes the

[1] Institute of Physical Geodesy, Darmstadt University of Technology, Petersenstrasse 13, 64287 Darmstadt, Germany. E-mail: roedelsperger@geod.tu-darmstadt.de

[2] Western Australian Centre for Geodesy and The Institute for Geoscience Research, Curtin University of Technology, GPO Box U1987, Perth, WA 6845, Australia.

[3] Now at: Asia-Pacific Applied Science Associates, PO Box 7650, Cloisters Square, Perth, WA 6850, Australia.

ocean water to expand/contract without altering its mass (e.g., MEIER and WAHR, 2002; ANTONOV *et al.*, 2002).

The Intergovernmental Panel of Climate Change (IPCC) estimated the sea-level rise in the 20th century to be 1.7 ± 0.5 mm/year, based on analyses obtained by tide gauge observations (CHURCH *et al.*, 2004; HOLGATE and WOODWORTH, 2004; CHURCH and WHITE, 2006; BINDOFF *et al.*, 2007). CAZENAVE and NEREM (2004) determined an accelerated rate of 3.1 mm/year between 1993 and 2003 by satellite altimetry. Both estimates include the combined influences of steric and non-steric effects.

Several studies conclude that in the second half of the last century the thermal expansion accounts for about 25% of the 20th century global average sea-level rise given by the IPCC (see Table 5.2 in BINDOFF *et al.*, 2007). ANTONOV *et al.* (2005) estimate the thermosteric effect between 1955 and 2003 based on *in situ* observations for the upper 700 m at 0.33 ± 0.04 mm/year. For the same time period and depth range, ISHII *et al.* (2006) provide an estimate of 0.36 ± 0.07 mm/year. Studies over the last decade, instead, show an accelerated thermosteric sea-level change becoming one of the major contributors to contemporary sea-level change as observed by satellite altimetry. ANTONOV *et al.* (2005) and ISHII *et al.* (2006) estimate a rate of 1.2 ± 0.5 mm/year for the upper 700 m while WILLIS *et al.* (2004) and LOMBARD *et al.* (2006) provide for the same time-period and depth range estimates of 1.6 ± 0.5 mm/year and 1.8 ± 0.4 mm/year, respectively. The latter estimates account for more than 50% of the total sea-level rise as observed by satellite altimetry (CAZENAVE and NEREM, 2004).

For projections on future sea-level changes for the 21st century, diverse climate modelling outcomes are available. The fourth assessment report of the IPCC gives an overview of the results until 2007 (MEEHL *et al.*, 2007). Depending on the used emission scenario, the projected sea-level rise is between 0.18 and 0.59 m with the thermosteric contribution assumed 70 to 75%.

It is well known that a net ocean-mass change (e.g., due to melting ice) causes a change in the Earth's gravity field (e.g., FARRELL and CLARK, 1967; KUHN *et al.*, submitted). It is not well known, however, that the thermosteric sea-level change also causes changes in the gravity field. In this study we investigate the sea-level rise caused by the thermosteric (temperature induced) and halosteric effect (salinity induced) over the next 2000 years, using the output of two climate model runs of the UVic ESCM under different CO_2 scenarios. Based on the steric sea-level change, we quantify the impact on the Earth's gravity field and show that, despite the fact that no mass is added or removed, the vertical redistribution of ocean watermasses (expansion or contraction) leads to changes in Earth's gravity field.

2. Climate Change Scenarios

The climate model used for this study is the Earth System Climate Model developed at the University of Victoria, Canada (UVic ESCM) version 2.6 and 2.7. The resolution

of the OGCM is 3.6° zonal by 1.8° meridional with 19 vertical layers with a maximum depth of 5396 m. The UVic ESCM 2.6 includes a three-dimensional spherical Ocean General Circulation Model (OGCM) coupled with a thermodynamic-dynamic sea-ice model, an energy-moisture balance atmospheric model and a land-ice model (WEAVER *et al.*, 2001). Land-ice sheets are not included in this version. Version 2.7 additionally includes a vegetation model, ocean biology and land surface processes (MAKARYNSKYY *et al.*, 2007). Land-ice sheets are included but do not melt and thus act mainly as large coolers.

The computation of two CO_2 emission scenarios for a period of 2000 years (2001–4001) was performed at the Western Australian Centre for Geodesy (MAKARYNSKYY *et al.*, 2005). The first scenario, computed with the UVic ESCM 2.6, simulates a CO_2 increase of 1% per year (3.65 ppm/year), starting from the initial concentration of 365 ppm, which is the observed CO_2 concentration in 1998 (KEELING and WHORF, 2005). Due to the continuous CO_2 increase in the atmosphere, this scenario reaches a level of its concentration, which is impossible to achieve even when burning all present day stored CO_2. However, this provides a good opportunity to study the behavior of the ocean due to intense heating and thus provides maximum effects. Under this scenario, sea ice melts almost completely and the total snow volume, accumulated on land and sea ice, reduces by about 40% of its initial amount.

In the second scenario, computed with version 2.7 of the UVic ESCM, CO_2 increases, starting again from 365 ppm, until it reaches the level of 3011 ppm (after 423 years). Afterwards it remains constant. The CO_2 emission rate during the first 423 years accelerates with time (6.3 ppm/year on average).

Due to the rigid-lid approximation, sea-level changes are not modelled explicitly. Mass-induced sea-level changes caused by mass transports within the oceans can be assessed by means of the surface water pressure under the lid (e.g., MAKARYNSKYY *et al.*, 2006). This study, however, only focuses on the steric sea-level change rather than mass-induced changes.

3. Steric Sea-Level Change

The steric effect $\Delta h_{z_1}^{z_2}$ between height z_1 and z_2 is defined as the difference of steric height $h_{z_1}^{z_2}$ of a specific water column with temperature T_0 and salinity S_0 and the water column with temperature $T_0 + \Delta T$ and salinity $S_0 + \Delta S$ (e.g., LANDERER *et al.*, 2005):

$$\Delta h_{z_1}^{z_2} = h_{z_1}^{z_2}(T_0 + \Delta T, S_0 + \Delta S) - h_{z_1}^{z_2}(T_0, S_0). \tag{1}$$

The steric height is the height difference between the water column of a specified density $\rho(T,S,p)$ and the water column of an ideal density (i.e., temperature of 0°C, salinity of 35 psu) $\rho(0,35,p)$,

$$h_{z_1}^{z_2} = \int_{z_1}^{z_2} \frac{\rho(T,S,p) - \rho(0,35,p)}{\rho(0,35,p)}\, dz. \tag{2}$$

To estimate separately the contribution of temperature and salinity, the steric effect can be subdivided into the thermosteric and the halosteric effects (e.g., ANTONOV *et al.*, 2002; LANDERER *et al.*, 2005). To compute the thermosteric effect, salinity is set constant and to compute the halosteric effect, the temperature is set constant in eq. (1).

Due to the increasing CO_2 concentration, the globally averaged surface air temperature rises from 13.5°C to 27.2°C between 2001 and 4001 in the first scenario. The global mean ocean temperature, averaged over all vertical layers, increases from 3.7 to 11.6°C. Melting sea ice and snow results in a global average decline of salinity from 34.748 to 34.694 psu.

Due to the warming and freshening of the ocean and the resulting expansion of the water, the global mean sea level rises by 5.2 m in 2000 years (Fig. 1). In terms of a global average, 97% (5.0 m in 2000 years) of the sea-level rise is caused by the thermosteric effect, whereas the halosteric effect instead, accounts only for 3% (0.2 m in 2000 years) and is positive due to decreasing salinity, thus amplifies the thermosteric effect. However, the halosteric effect can become more dominant in Polar Regions (see below).

In scenario 2 between 2001 and 4001, the globally averaged surface air temperature increases from 14.0 to 24.8°C and the globally averaged ocean temperature rises from 5.1

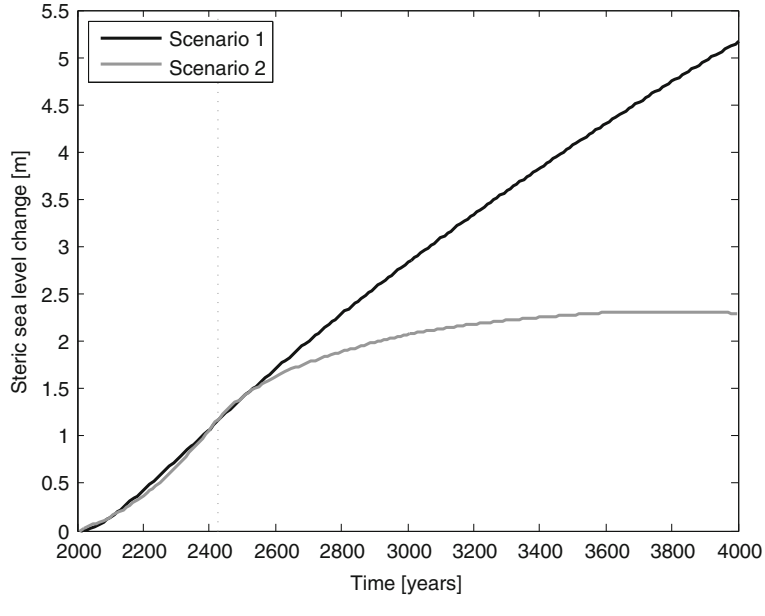


Figure 1
Globally averaged steric sea level change in scenario 1 (black line) and scenario 2 (grey line). The dashed line indicates the halt of CO_2 increase in scenario 2.

to 10.4°C. Temperature and salinity in the year 2001 are different to scenario 1, because the default initial settings in version 2.7 are different to version 2.6. While the surface air temperature and the temperature in the surface layer of the ocean directly respond (within a few decades) to the halt of CO_2 increase after 423 years, the ocean temperature in the deeper layers still rises during several hundred years more due to the inertial nature of the oceanic processes. Globally averaged salinity in this scenario rises from 34.864 to 34.885 psu, even though great amounts of snow and ice melt. This is not yet understood but may be caused by numerical problems in the salinity fluxes during the computation (Michael Eby, personal communication). In this scenario, the globally averaged sea-level rises by 2.3 m due to the steric effect (Fig. 1). That is almost twice the sea-level rise, occurring during the time of CO_2 increase (first 423 years), which is due to the warming of the deeper ocean layers that continues well beyond the time of CO_2 increase.

In the first 100 years, both scenarios did not differ significantly in global average with a steric sea-level rise of 0.17 m. The regional steric effect, however, varies in scenario 1 between −0.03 and 0.46 m and in scenario 2 between −0.68 and 1.11 m. Figure 2 shows the thermosteric and halosteric sea-level change of both scenarios. The strong rising sea level in scenario 1 in parts of the Atlantic Ocean and south of Africa is in areas where the heat penetrates rapidly into deeper layers (e.g., Landerer et al., 2005). Between 2001 and 4001, the regional steric sea-level change in scenario 1 ranges from 0.0 to 8.2 m and is strongly correlated with the sea bottom topography, e.g., the Mid-Atlantic Ridge can be clearly identified (Fig. 3). This is expected because due to the larger water masses involved, the warming of deep water columns causes a higher steric effect than the warming of shallow water columns.

In mid and low latitudes, the steric effect is mainly caused by warming, whereas in the Arctic Ocean, where the ocean warming is less, between 40 and 90% of the total effect is caused by decreasing salinity. The change of density due to salinity changes is higher in colder water and thus the importance of salinity is growing for lower temperatures, thus in the Polar Regions. This shows that the regional contribution of salinity can be very important, which has already been observed using *in situ* data (e.g., Antonov *et al*., 2002; Ishii *et al*., 2006).

In scenario 2, the regional steric effect between 2001 and 4001 varies spatially between −0.1 and 3.8 m (Fig. 3). The correlation with topography is not as noticeable as in scenario 1, because the heat does not penetrate as deep as in scenario 1. In scenario 1, the strongest sea-level rise takes place in the Pacific Ocean; in scenario 2, the Atlantic Ocean sea level increases most.

4. Gravitational Attraction and Potential

The Earth's gravitational field, which is produced by the Earth's internal mass distribution, reacts to any rearrangement of these masses. In case of the steric effect, there is no mass change but a rearrangement due to expansion/contraction of seawater that

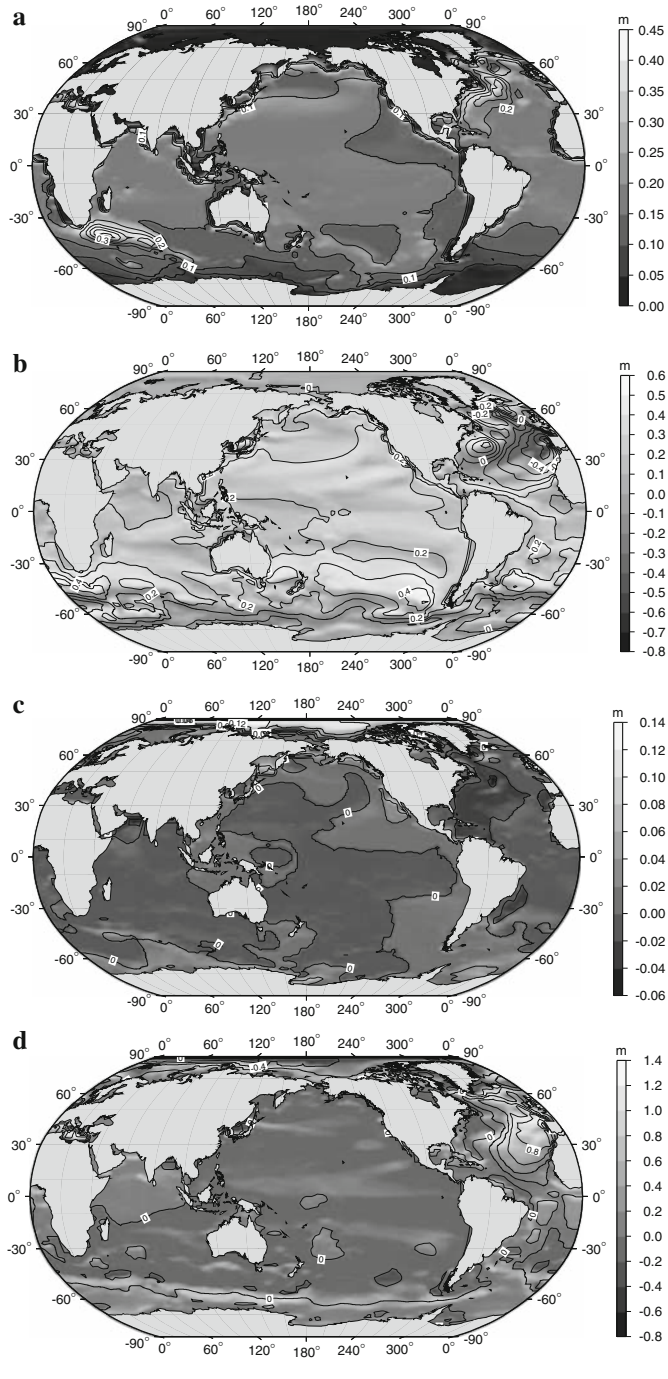

a
m
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
b
m
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
c
m
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
d
m
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8

Figure 2
Steric sea-level change between 2001 and 2101; (a) thermosteric effect in scenario 1; (b) thermosteric effect in scenario 2; (c) halosteric effect in scenario 1; (d) halosteric effect in scenario 2.

mainly acts in a vertical direction. To obtain the gravitational effect caused by the steric sea-level change here, the ocean surface in year 2001 is considered an equipotential surface, e.g., the geoid. The expansion of ocean water masses due to the steric effect causes spatially different vertical movement of the ocean surface and thus any observation station, especially close to the expanding/contracting water masses, will observe changes in gravitational attraction and potential.

In this study, the observation stations (e.g., computation points) are located on the initial ocean surface, thus closest to the changing water masses. Therefore, the steric effect will place water masses above the observation stations, thus the effect on gravity will be negative. The steric effect is not uniform across the globe (Figs. 2–3), especially due to the uneven distribution of continental landmasses and ocean depth. Thus, the gravitational effect will differ as well and the assumed observation stations will no longer be located on the initial ocean surface.

The vertical shift of the initial equipotential surface (e.g., geoid) ΔN can be derived by the theorem of Bruns (e.g., TORGE, 2003)

$$\Delta N = \frac{\Delta V}{g} \quad \text{with} \quad \Delta V = V_1 - V_0, \tag{3}$$

with V_0 being the gravitational potential induced by the initial ocean water mass distribution. V_1 is the gravitational potential of the ocean after its mass has been redistributed by the steric effect and g is a mean gravity value (e.g., 9.81 $\mathrm{m} \cdot \mathrm{s}^{-2}$).

Newton's law of gravitation (e.g., TORGE, 2003; HEISKANEN and MORITZ, 1967) can derive the gravitational potential and attraction of any mass distribution. The gravitational potential is defined as a volume integral over the complete volume v of the masses considered

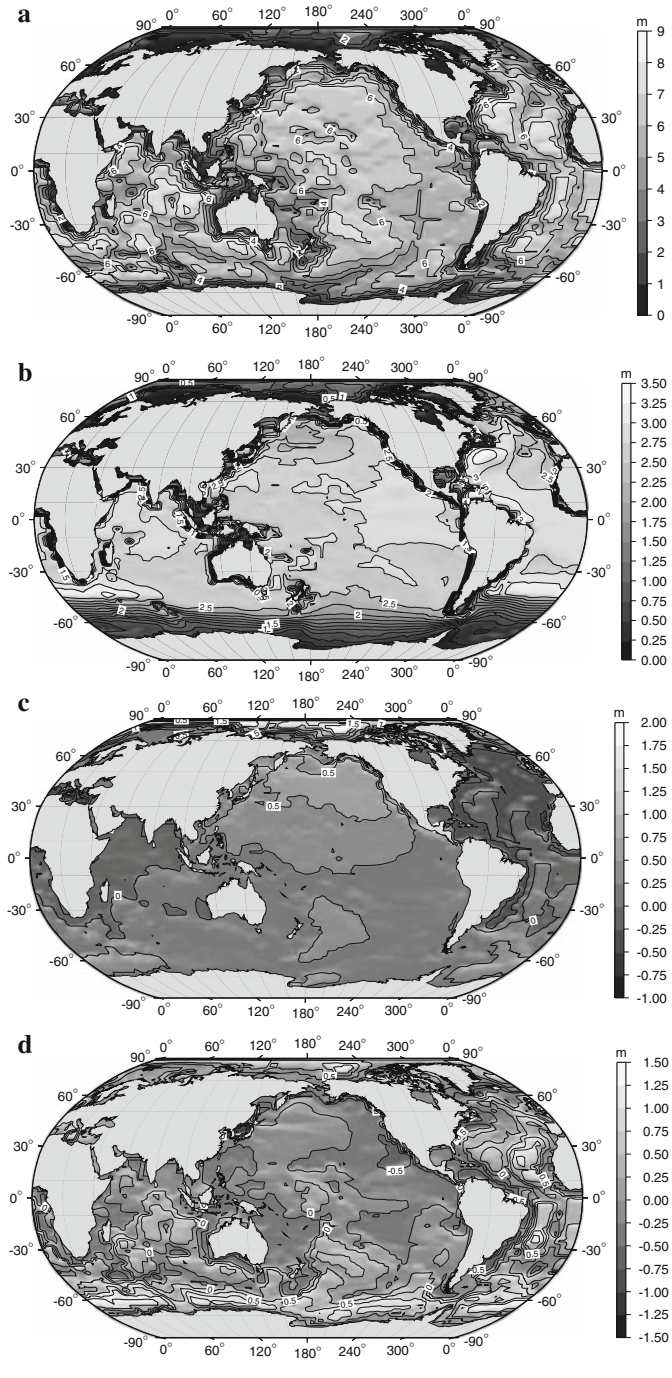
$$V = G \iiint_v \frac{1}{l} \rho \, dv', \tag{4}$$

whereas G is the gravitational constant, ρ the density and l the distance between computation point and volume element dv'. The gravitational attraction is given by

$$\vec{b} = \text{grad } V = \left(\frac{\partial V}{\partial X}, \frac{\partial V}{\partial Y}, \frac{\partial V}{\partial Z} \right). \tag{5}$$

It is possible to solve the integral given by eq. (4) for elementary bodies such as point masses, spherical shells and prisms.

In a very simplistic model, the ocean water masses can be considered as a spherical shell with constant density ρ_0. Due to the steric effect the shell expands and the new density ρ_1 is given in spherical approximation by


a
m
9
8
7
6
5
4
3
2
1
0
b
m
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
c
m
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
d
m
1.50
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
-1.25
-1.50

◀ Figure 3
Steric sea-level change between 2001 and 4001; (a) thermosteric effect in scenario 1; (b) thermosteric effect in scenario 2; (c) halosteric effect in scenario 1; (d) halosteric effect in scenario 2.

$$\rho_1 = \rho_0 \cdot \frac{R^3 - (R-d)^3}{(R+h)^3 - (R-d)^3}, \tag{6}$$

whereas R is the mean radius of Earth, d the depth of the ocean and h the steric sea-level rise.

The effect of the steric sea-level change on gravitational potential and attraction can be calculated by subtracting the potential and attraction of the initial ocean from the potential and attraction of the expanded ocean. For a spherical shell, the effect on gravitational potential ΔV can be derived from eq. (4)

$$\Delta V = V_1 - V_0 \; = \pi G \cdot \left[2\, \rho_1\, h \cdot (2R + h) + (\rho_1 - \rho_0) \cdot \frac{d}{R} \cdot \left((2R - d)^2 + \frac{1}{3} d^2 \right) \right]. \tag{7}$$

The effect on gravitational attraction in vertical direction Δb follows from eq. (5) by

$$\Delta b = b_1 - b_0 = \pi G \cdot (\rho_1 - \rho_0) \cdot \frac{d}{R^2} \cdot \left((2R - d)^2 + \frac{1}{3} d^2 \right), \tag{8}$$

with b_0 being the effect on gravitational attraction induced by the initial ocean water mass distribution and b_1 that of the ocean water masses changed by the steric effect.

Initial insight in to the magnitude of the corresponding effects can be obtained by applying a spherical shell only (cf. eqs. 7 and 8). This very simplistic model neglects continents and variations of the ocean depth. Applying eqs. (7) and (8) with radius R = 6378137 m, density ρ_0 = 1030 kg $\cdot$ m^{-3}, depth d = 5000 m and a steric sea-level change of h = 5 m, which corresponds to the global average steric effect in scenario 1 in 2000 years, provides the estimates

$$\Delta V = -1.078 \cdot 10^{-5} \mathrm{m}^2 \cdot \mathrm{s}^{-2}, \tag{9}$$

$$\Delta b = -0.4314 \cdot 10^{-5} \mathrm{m} \cdot \mathrm{s}^{-2}. \tag{10}$$

With g = 9.81 m $\cdot$ s^{-2}, the resulting shift ΔN in the geoid can be calculated using the theorem of Bruns given by eq. (3)

$$\Delta N = 1.1\ \mu\mathrm{m}. \tag{11}$$

This shift is very small, especially when considering the time (2000 years) in which this effect occurs. In case all ice would melt, the shift of an equipotential surface can be almost 100 m (KUHN *et al.*, submitted) and more than one centimeter for the cryospheric ice-mass decline over the period of 2002 to 2007 (e.g., BAUR *et al.*, submitted).

To obtain a spatial distribution of the effect, the new sea surface, given in the form of a Digital Elevation Model (DEM) and a Digital Density Model (DDM), can be

subdivided into elementary bodies. The DEM and DDM hold the sea surface height and density information, respectively, in a regular two-dimensional geographic grid. The most reasonable elementary bodies for a geographic coordinate system are spherical tesseroids. A spherical tesseroid (Fig. 4) is a body bounded by two concentric spheres defined by radius r_1 and r_2, two coaxial cones defined by colatitudes ϑ_1 and ϑ_2 and two meridional planes defined by longitudes λ_1 and λ_2 (ANDERSON, 1976).

The gravitational potential V in P(ϑ, λ, r) of a tesseroid follows from eq. (4) (e.g., HECK and SEITZ, 2006)

$$V = G\rho \cdot \int_{\lambda'=\lambda_1}^{\lambda_2} \int_{\vartheta'=\vartheta_1}^{\vartheta_2} \int_{r'=r_1}^{r_2} \frac{1}{l}\, dv, \tag{12}$$

with

$$dv = r'^2 \sin\vartheta' dr' d\vartheta' d\lambda', \tag{13}$$

$$l = \sqrt{r^2 + r'^2 - 2rr'(\cos\vartheta\cos\vartheta' + \sin\vartheta\sin\vartheta'\cos(\lambda' - \lambda))}. \tag{14}$$

This is an ellipsoidal integral and thus analytically not solvable. Different approaches can be used to solve it (e.g., HECK and SEITZ, 2006; KUHN, 2000; ANDERSON, 1976):

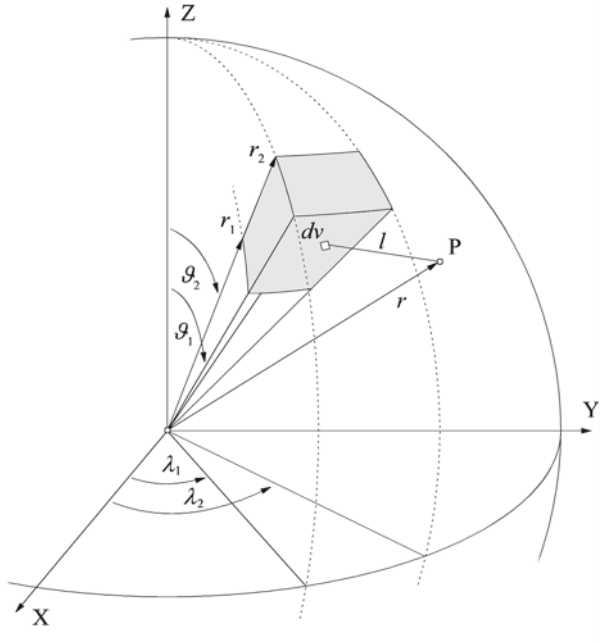


Figure 4
Definition of a tesseroid.

- Numerical integration
- Substitution of tesseroids by prisms having the same volume, density and height as the tesseroids
- Analytical solution for a computation point P located at the polar axis of the geographic coordinate system.

Here, only the latter two approaches are considered.

4.1. Substitution of Tesseroids by Mass Equal Prisms

Based on the following relations (e.g., ANDERSON, 1976; KUHN, 2000, 2003)

$$m_{\text{tesseroid}} = m_{\text{prism}}, \quad \Delta z = \Delta r = r_2 - r_1, \tag{15}$$

the dimensions Δx, Δy and Δz of the mass equal prism are obtained by

$$\Delta x = r_m \, \Delta\vartheta, \quad \Delta y = r_m \, \cos\vartheta_m \, \Delta\lambda, \quad \Delta z = \Delta r. \tag{16}$$

with

$$r_m = \frac{r_1 + r_2}{2} \quad \text{and} \quad \vartheta_m = \frac{\vartheta_1 + \vartheta_2}{2}. \tag{17}$$

Here Δx, Δy, and Δz represent coordinate differences given with respect to a local topocentric (3D Cartesian) coordinate system where the z-axis coincides with the local zenith direction going through the center of the tesseroid.

For the gravitational potential of a homogeneous rectangular prism in a local topocentric coordinate system (x, y, z) for a computation point P, located in the origin of the coordinate system, eq. (4) transforms to (e.g., NAGY *et al.*, 2000)

$$V = G\rho \cdot \int_{x=x_1}^{x_2} \int_{y=y_1}^{y_2} \int_{z=z_1}^{z_2} \frac{1}{l} dv, \tag{18}$$

with

$$dv = dx\,dy\,dz, \quad l = \sqrt{x^2 + y^2 + z^2}. \tag{19}$$

The integration of eq. (18) can be performed analytically and delivers a closed formula for the gravitational potential

$$\begin{aligned} V = G\rho \cdot \Big| \Big| \; & xy \ln(z + l) + yz \ln(x + l) + zx \ln(y + l) \\ & - \frac{x^2}{2} \arctan\frac{yz}{xl} - \frac{y^2}{2} \arctan\frac{zx}{yl} - \frac{z^2}{2} \arctan\frac{xy}{zl} \Big|_{x=x_1}^{x_2} \Big|_{y=y_1}^{y_2} \Big|_{z=z_1}^{z_2}, \end{aligned} \tag{20}$$

and for the vertical component of the gravitational attraction

$$b = G\rho \cdot |\,|\,| \; x\ln(y+l) + y\ln(x+l) - z\arctan\frac{xy}{zl}\Big|_{x=x_1}^{x_2}\Big|_{y=y_1}^{y_2}\Big|_{z=z_1}^{z_2}. \tag{21}$$

4.2. Analytical Integration for a Computation Point at the Polar Axis

A closed formula for gravitational potential and attraction of tesseroids can be derived from eq. (12) for a computation point P located at the polar axis (Z axis) of the geographic coordinate system with $\vartheta = 0°$ (HECK and SEITZ, 2006). In this case, the distance l given by eq. (14) simplifies to

$$l = \sqrt{r^2 + r'^2 - 2rr'\cos\vartheta'}. \tag{22}$$

To calculate gravitational potential and attraction at any arbitrary location on the sphere it is necessary to introduce a local geographic coordinate system (essentially a local polar coordinate system on the sphere) with the local polar axis being coincident with the local zenith direction going through the computation point P (Fig. 5). Since global DEM data are usually given with respect to a geographic coordinate system, a transformation has to be performed for every combination between computation point and mass element. The transformation of a mass element's location $P'(\psi', \alpha')$ given with respect to a local geographic coordinate system into the corresponding location $P(\varphi, \lambda)$ of the global geographic coordinate system is given by

$$\cos\vartheta' = \cos\vartheta\cos\psi' + \sin\vartheta \;\; \sin\psi'\cos\alpha', \tag{23}$$

$$\tan(\lambda' - \lambda) = \frac{\sin\alpha'}{\cot\psi'\sin\vartheta - \cos\vartheta\cos\alpha'}. \tag{24}$$

Utilizing the relations in eqs. (23) and (24) the steric sea-level change at point $P'(\varphi', \lambda')$ can be determined from a global grid given in geographic coordinates, e.g., through interpolation. In spite of the high numerical effort, the results obtained by this method are numerically accurate because no approximation error is made due to the use of closed analytical formulas.

4.3. Computation of the Impact of the Steric Effect on the Gravity Field

We developed different FORTRAN codes for the computation of the effect on the gravitational potential and attraction based on diverse elementary bodies (KUHN, 2000, 2003). The representation of the water masses is done with respect to a geographic coordinate system by two DEMs containing the height of the lower and upper bound of the water masses and one DDM containing the lateral variable density. Optionally, several topographic grids with different spatial resolutions (finest resolution is closest to the computation point) can be used in order to speed up the global integration.

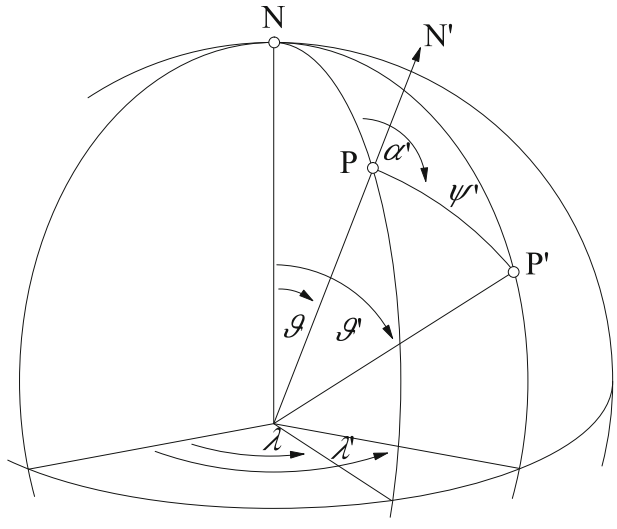


Figure 5
Relation between the (global) geographic coordinate system (λ, ϑ) and the local geographic coordinate system (α', ψ'). Here, the north pole is indicated by N, whereas the local zenith direction through point P is indicated by N′.

Here, the effect on gravitational potential and attraction is the difference between two separate calculations (cf. eq. (3): $\Delta V = V_1 - V_0$). The computation of the effect of the initial ocean is done by setting all elements of the DEM of the upper bound to 0 m, the elements of the DEM of the lower bound to the bathymetric depth given by UVic ESCM and all elements of the DDM to a density of 1030 kg $\cdot$ m^{-3}. Although the OGCM provides variable density information for the initial ocean, the choice is of minor importance as we only deal with difference values. For the expanded ocean, the upper bound is represented by the steric height and the density of each element is calculated with eq. (6), whereas the lower bound remains the same corresponding to the sea bottom topography.

To quantify the approximation error made by using prisms we compute the gravitational attraction and potential of a homogeneous spherical shell represented by global DEMs and DDMs of different spatial resolutions but of constant height and density. The result is compared to the analytic solution for the spherical shell (cf. eqs. (9) and (10)). The results in Table 1 were computed using a spherical shell with radius R = 6378137 m, density ρ_0 = 1030 kg $\cdot$ m^{-3}, depth d = 5000 m and a steric sea-level change of h = 5 m.

As can be seen in Table 1, the relative error in g_0 (initial ocean) and g_1 (expanded ocean) as compared to the reference value for resolutions ranging between 3.6° $\times$ 1.8° and 0.4° $\times$ 0.4° lies between $4.3 \cdot 10^{-5}$ m s^{-2} (1.00%) and $0.5 \cdot 10^{-5}$ m s^{-2} (0.12%). The approximation error is getting smaller with higher resolutions. Here, only the difference $\Delta g = g_1 - g_0$ is of interest, which shows a much smaller relative error ($\leq$0.05%) as

Table 1

Gravitational attraction and potential of the initial ocean (g_0/V_0, first column) and the expanded ocean (g_1/V_1, second column), as well as the difference ($\Delta g/\Delta V$, third column) for a spherical shell and grids of different resolutions. The value for the spherical shell provides the reference value.

Gravitational attraction			
	g_0 [10^{-5} ms^{-2}]	g_1 [10^{-5} ms^{-2}]	Δg [10^{-5} ms^{-2}]
Spherical shell	431.4521	431.0208	−0.4314
3.6° × 1.8°	427.1725	426.7409	−0.4316
1.0° × 1.0°	429.9593	429.5278	−0.4315
0.4° × 0.4°	430.9473	430.5158	−0.4315
Gravitational potential			
	V_0 [m^{2}s^{-2}]	V_1 [m^{2}s^{-2}]	ΔV [10^{-5} m^{2}s^{-2}]
Spherical shell	27518.6087 8077	27518.6087 6999	−1.078
3.6° × 1.8°	27517.8543 1477	27517.8541 9584	−11.893
1.0° × 1.0°	27518.5597 3922	27518.5596 8963	−4.959
0.4° × 0.4°	27518.6084 9243	27518.6084 6736	−2.506
Three grids	27518.5491 3803	27518.5491 2437	−1.366

approximation errors made in g_0 and g_1 are very similar and thus almost cancel each other out. Generally, a computation with a resolution of 1.0° × 1.0° or finer provides an adequate accuracy for the effect of the steric sea-level change on gravitational attraction.

As the effect on gravitational potential is very small, numerical problems in the calculation using tesseroids substituted by mass equal prisms occur. The results for the potential calculated using quadruple precision are shown in Table 1, generally showing higher relative errors than for the gravitational attraction.

A better solution regarding accuracy and computation time can be found when using several grids with resolutions dependent on the distance, since the gravitational potential and attraction effects decrease rapidly with distance, thus a coarser resolution can be used for more distant masses. For three grids, arranged around the computation point as shown in Figure 6, the computation time is greatly reduced (about 1 min per point using a 256 MHz processor). The approximation error for V_0 and V_1 is higher than for a grid with a resolution of 1.0° × 1.0°, but the error for the difference ΔV is much smaller. The relative error made in ΔV is at most 25% for a depth of 5000 m and a steric effect of 5 m and gets smaller for smaller depths and smaller steric sea-level changes. This error level is considered sufficient to provide the general spatial behaviour of the effect on gravitational potential. However, it is prudent to treat the absolute magnitudes with care.

5. Results

In order to quantify the effect on gravitational potential and attraction due to steric sea-level variations obtained from scenario 1, we employed three different computation techniques. A rough approximation was done by calculating the effect through spherical shells (cf. eqs. (7) and (8)) (hereafter referred to as 'shell-method'). The dimension of the

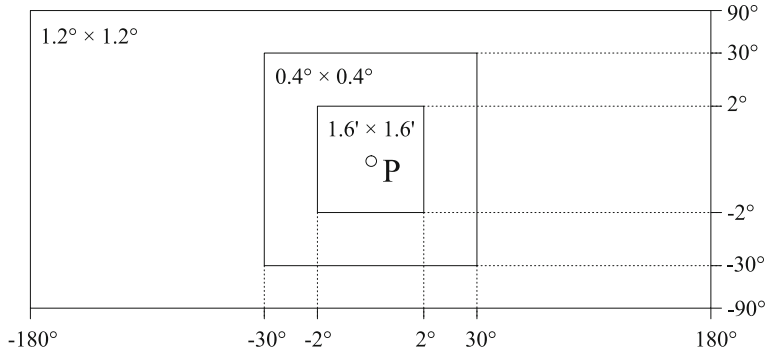


Figure 6
Configuration of three grids in different spatial resolutions.

spherical shell is defined by the depth and steric height of the ocean element directly beneath each computation point, thus neglecting the spatial variation of steric sea level around the computation point.

A more precise calculation was done for tesseroids expressed by mass equal prisms according to eqs. (20) and (21) (hereafter referred to as 'prism-method'). For the effect on gravitational potential, the grid configuration shown in Figure 6 was used. Steric height and ocean depth were interpolated from the original resolution of 3.6° × 1.8° given by UVic ESCM to the required resolutions using the cubic interpolation algorithm provided by Matlab (MATHWORKS, 2002). The density was then calculated applying eq. (6).

Furthermore, the effect on gravitational attraction and potential was calculated using the closed formulas for tesseroids given by eqs. (12), (13) and (22) (hereafter referred to as 'tesseroid-method'). As discussed in paragraph 4.2, the data (density and steric height), given in a global geographic coordinate system has to be transformed for each computation point into a local geographic coordinate system. Since the tesseroid formulas are analytical, a grid resolution of 1.8° × 0.9° in the local coordinate system is sufficient. The spherical coordinates of all elements in the local geographic coordinate system were transformed into the global geographic coordinate system using eqs. (23) and (24). Depth, steric height and density were then adopted from the nearest neighbour.

Figure 7(a) shows the spatial distribution of the effect of the steric sea-level change on gravitational attraction calculated with the prism-method in scenario 1. The corresponding steric sea-level change was given in Figure 3(a).

The effect on gravitational attraction ranges from 0.0 to $-0.7 \cdot 10^{-5}$ m s^{-2}. According to the steric effect always being positive, the effect on gravitational attraction is always negative (due to the expansion some water masses are located above the computation point). The relation between the effect on gravitational attraction and the steric sea-level change is practically linear. The main effect on gravitational attraction is caused by the

element near the computation point, and the influence decreases rapidly with distance. It is practically zero for elements located at a spherical distance of more than 5°. Thus, the effect over land is practically zero.

The dominating effect of the elements in the immediate vicinity of the computation point can be seen as well, when the effect on gravitational attraction is calculated with the shell method. The result is not shown here, since the figure is basically the same as Figure 7(a). The mean deviation between the shell-method and the prism-method for the effect on gravitational attraction is $\pm 5 \cdot 10^{-9}$ m s^{-2} with a maximum deviation of $7.0 \cdot 10^{-8}$ ms^{-2} located in the Arctic Ocean. There, the dimensions of the prisms are smallest and therefore the approximation with a spherical shell shows a higher approximation error. For the absolute value of gravitational attraction (g_0 or g_1), the error is much higher with a mean deviation of $\pm 72 \cdot 10^{-5}$ m s^{-2} and a maximum of $130 \cdot 10^{-5}$ m s^{-2}. Thus, for the absolute value, elements in greater distances have a greater impact.

Calculating the effect on gravitational attraction with the tesseroid-method shows a similar image as well. The difference between this method and the prism-method is with

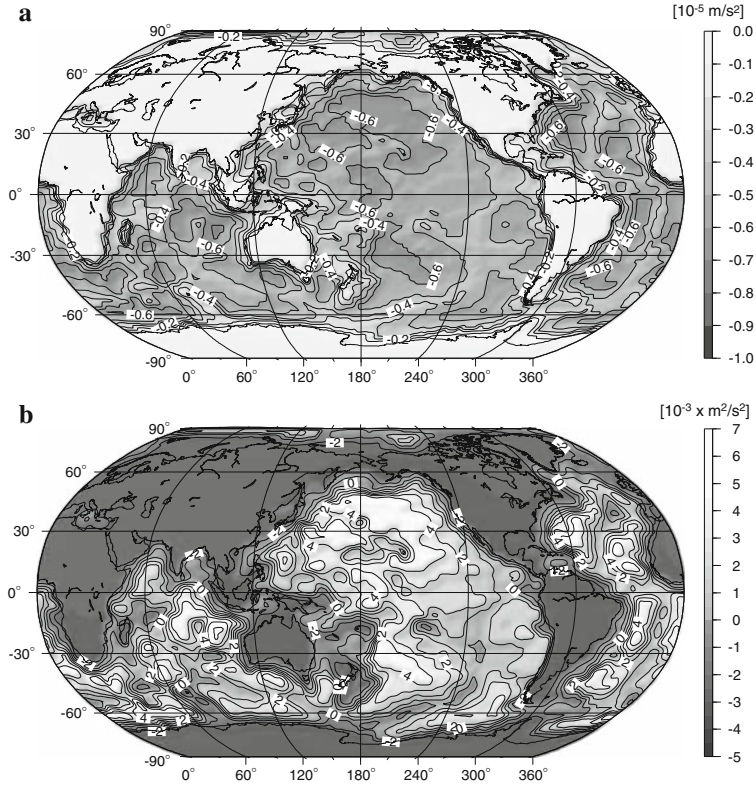


Figure 7
Effect of the steric sea-level change between years 2001 and 4001 on gravitational attraction (a) and gravitational potential (b).

a mean deviation of $\pm 5.7 \cdot 10^{-8}$ m s^{-2} and a maximum deviation of $8.0 \cdot 10^{-7}$ ms^{-2} about one order of magnitude higher than the difference between the prism- and the shell-method. However, for the single gravitational attraction of the water masses, the mean deviation is $\pm 23 \cdot 10^{-5}$ m s^{-2} and the maximum difference amounts to $52 \cdot 10^{-5}$ m s^{-2}. The differences are in part caused by the approximation error of the prism method, although probably mostly by mass changes due to the very simple interpolation of the DEM data (near-neighbor). Generally, the prism- and the shell-method show a good accordance in terms of the gravitational effect caused by the steric sea-level change. For the gravitational attraction of the water masses, the difference between the prism- and tesseroid-method is smaller.

The computation with the shell-method for the effect on gravitational potential gives values between 0.0 and $2.9 \cdot 10^{-5}$ m^{2}s^{-2} and the general behavior is similar to the effect on gravitational attraction. Using the tesseroid-method, the effect ranges from $-3.2 \cdot 10^{-3}$ m^{2}s^{-2} to $6.1 \cdot 10^{-3}$ m^{2}s^{-2} (Fig. 7(b)). Generally, the effect on gravitational potential is negative in continents or shallow water regions and positive in deep ocean regions, with the highest effect coincident with the highest steric sea-level change. Thus, the values are not only two orders of magnitude larger than the values obtained by the shell-method, but also have different signs. The prism-method delivers a similar distribution. The mean deviation between the prism- and the tesseroid-method is $1.2 \cdot 10^{-4}$ m^{2}s^{-2} with a maximum difference of $7.9 \cdot 10^{-4}$ m^{2}s^{-2}.

The reason for the different magnitude between the value obtained from the shell-method and the prism-method can be found by looking at the effect of a single small element. Figure 8 shows the contribution of a small tesseroid with constant mass for the expansion of a spherical shell in function of the distance from the computation point. For a very small area around the computation point, the effect of the expansion on gravitational potential is positive. Then a minimum with negative value is reached and the effect converges to zero. For a homogeneous spherical shell, the contribution of elements with a negative effect exceeds slightly. For an asymmetrical expansion, mostly caused by the irregular distribution of land-masses, the effect on gravitational potential can as well be positive and much larger.

In general, this shows that the effect on gravitational potential is not only dependent on the magnitude of the sea-level change, but also highly dependent on the variability of the steric effect in its immediate vicinity. An outstanding large steric change causes a high positive effect on potential and an outstanding low steric change causes a high negative effect on gravitational potential. This can be seen at the continents near the coast, where the steric effect at the computation point is zero and nearby sea-level change leads to a high negative potential difference. Therefore, for instance, Australia and islands will undergo a stronger negative change in the potential difference than huge land-masses like Europe and Asia.

Using Bruns theorem given by eq. (3) to calculate the effect on geoid heights delivers values ranging between -0.3 and 0.6 mm, which is considerably higher than the effect estimated with the spherical shell (cf. eq. 11), but still very small.

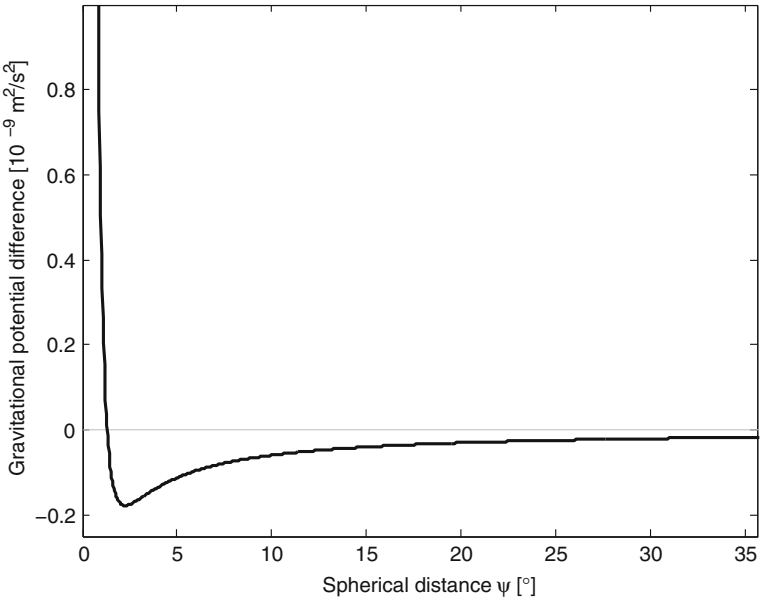


Figure 8
Contribution of a small element as a function of the spherical distance from the computation point to the effect on gravitational potential of the expansion of a homogeneous spherical shell.

6. Conclusion

This study analysed two climate model runs, performed with the UVic ESCM, in terms of long-term changes of ocean temperature, salinity and steric sea level and the impact of the latter on the Earth's gravity field. They do not represent a prediction of the future, since both should be considered as worst-case or even impossible scenarios (especially scenario 1).

Over 2000 years, the steric effect causes a globally averaged sea-level rise of more than 5 m and 2 m in scenario 1 and 2, respectively. The highest sea-level rise occurs in the open and deep ocean. On a global average, about 97% of this sea-level rise is induced by temperature changes. During the first 100 years, the globally averaged steric sea-level rise is 0.17 m in both scenarios, which is in agreement with the fourth IPCC report (MEEHL *et al.*, 2007), although at the lower end of the projected sea-level changes.

For scenario 1, the impact of the steric effect on the gravity field was computed. The change of gravitational attraction ranges from 0.0 to $-0.7{\cdot}10^{-5}$ m s^{-2}. The relationship of steric sea-level change and gravitational attraction is almost linear with a factor of $-0.086 \cdot 10^{-5}$ ms^{-2} per 1 meter steric sea-level rise. While these effects are considerable compared to other geophysical effects that have smaller signals, the change in geoid height is rather small and ranges between -0.3 and 0.6 mm.

For the computation of gravitational potential, the tesseroid-method is very efficient and the calculation is much faster than the prism-method. However, in this case a very simple interpolation without mass conservation was used. For a more accurate interpolation, the computation time will increase considerably. The shell-method shows a very good accordance with the prism-method in the effect on gravitational attraction but cannot be used for the calculation of the effect on gravitational potential.

Since the steric sea-level change will be accompanied by mass changes due to water exchange with landice and other terrestrial water storages, these effects will exceed the effect on the gravitational potential and attraction due to steric effect in magnitude several times (e.g., MAKARYNSKYY *et al*., 2006; KUHN *et al*., submitted; BAUR *et al*., submitted). For example Kuhn *et al*. (submitted) have shown that the maximum shift of an equipotential surface can be almost 100 m if all currently land-based ice masses melt. For cryospheric melting over the last 5 years BAUR *et al*. (submitted) show a maximum shift of more than 1 centimeter.

The effects shown here have to be considered as maximum effects and are assumed to occur over the rather long time period of 2000 years. Therefore, for the steric effect expected in the next few hundred years the gravitational and potential effect will be much smaller and can be safely neglected for the gravitational potential.

To the authors' best knowledge, this study shows for the first time numerical evidence that the gravitational effects caused by the steric effect are indeed very small but not zero. The effects on gravitational attraction cannot be sensed by satellite missions such as CHAMP, GRACE and in the future GOCE, but could be sensed with precise, absolute gravimeters.

Acknowledgements

This study was funded by Australian Research Council (ARC) Discovery Project grant DP0345583 and financially supported by the Frauenförderung of Darmstadt University of Technology, Germany. This is TIGeR Publication number 106.

REFERENCES

ANDERSON, E.G. (1976), *The effect of topography on solutions of Stokes' problem*, Unisurv. Report S.14, School of Surveying and Spatial Information Systems, University of New South Wales, Australia, 252 pages.

ANTONOV, J.I., LEVITUS, S., and BOYER, T.P. (2002), *Steric sea-level variations during 1957–1994: Importance of salinity*, J. Geophys. Res. 107 (C12). doi:10.1029/2001JC000964.

ANTONOV, J.I., Levitus, S., and BOYER, T.P. (2005), *Thermosteric sea level rise, 1955–2003*, Geophys. Res. Lett. 32, L12602. doi:10.1029/2005GL023112.

BAUR, O., KUHN, M., and FEATHERSTONE, W.E. (submitted), GRACE-derived ice-mass variations and their effect on global sea-level change patterns. J. Geophys. res.

BINDOFF, N.L., WILLEBRAND, J., ARTALE, V., CAZENAVE, A., GREGORY, J., GULEV, S., HANAWA, K., LE QUÉRÉ, C., LEVITUS, S., NOJIRI, Y., SHUM, C.K., TALLEY, L.D., UNNIKRISHNAN, A. (2007), Observations: Oceanic climate

change and sea level, In *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller, eds.) (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA).

CAZENAVE, A., and Nerem, R.S. (2004), *Present-day sea level change: Observations and causes*, Rev. Geophys. 42, RG3001. doi: 10.1029/2003RG000139.

CHURCH, J.A., WHITE, N.J., COLEMAN, R., LAMBECK, K., and MITROVICA, J.X. (2004), *Estimates of the regional distribution of sea-level rise over the 1950 to 2000 period*, J. Clim. 17 (13), 2609–2625.

CHURCH, J.A., and WHITE, N.J. (2006), *A 20th century acceleration in global sea-level rise*, Geophys. Res. Lett. 33, L01602. doi: 10.1029/2005GL024826.

DOUGLAS, C.D., and PELTIER, W.R. (2002), *The puzzle of global sea-level rise*, Physics Today 55 (3), 35–40.

FARRELL, W.E., and CLARK, J.A. (1976), *On postglacial sea level*, Geophys. J. R. astr. Soc. *46*, 647–667.

HECK, B., and SEITZ, K. (2006), *A comparison of the tesseroid, prism and point-mass approaches for mass reductions in gravity field modeling*, J. Geodesy. doi: 10.1007/S00190–006-0094-0.

HEISKANEN, W.A., and MORITZ, H., *Physical Geodesy* (W.H. Freeman and Company, USA, 1967), 364 pages.

HOLGATE, S.J., and WOODWORTH, P.L. (2004), *Evidence for enhanced coastal sea-level rise during the 1990s*, Geophys. Res. Lett. 31, L07305. doi: 10.1029/2004GL019626.

HOUGHTON, J. T., *Global Warming, The Complete Briefing*, Third Edition (University Press, Cambridge, 2004).

ISHII, M., KIMOTO, M., SAKAMOTO, K., and IWASAKI, S. (2006), *Steric sea-level changes estimated from historical ocean subsurface temperature and salinity analysis*, J. Oceanog. *62* (2), 155–170.

KEELING, C.D., and WHORF, T.P. (2005), *Atmospheric CO_2 records from sites in the SIO air sampling network*. In: *Trends: A Compendium of Data on Global Change*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., USA.

KUHN, M. (2000), *Geoidbestimmung unter Verwendung verschiedener Dichtehypothesen*. Deutsche Geodätische Kommission, Reihe C, Heft 520, München.

KUHN, M. (2003), *Geoid determination with density hypothesis from isostatic models and geological information*, J. Geodesy 77, 50–65.

KUHN, M., FEATHERSTONE, W.E., MAKARYNSKYY, O., and KELLER, W., *Changes in the Earth's gravity field, global sea-level and land-ocean distribution from a simulated total melt of polar ice sheets*, Global Planet Change, submitted.

LANDERER, F.W., JUNGCLAUS, J.H., and MAROTZKE, J. (2005), *Regional dynamic and steric sea-level change in response to the IPCC-A1B scenario*, J. Phys. Oceanog., in press.

LOMBARD, A. *et al*. (2006), *Perspectives on present-day sea-level change: A tribute to Christian le Provost*, Ocean Dyn. 56 (5–6). doi: 10.1007/ s10236-005-0046-x.

MAKARYNSKYY, O., KUHN, M., and FEATHERSTONE, W.E. (2005), Modelling future sea-level change under greenhouse warming scenarios with an Earth System Model of intermediate complexity. In Gravity, Geoid and Space Missions Jekeli, C., Bastos, L., Fernandez J., (eds.), . IAG Symposia 129 (Springer Berlin, Heidelberg, New York, 2005) pp. 260–265.

MAKARYNSKYY, O., KUHN, M., EATHERSTONE W.E. (2007), *Long-term sea level projections with two versions of a global climate model of intermediate complexity and corresponding changes to the Earth's gravity field*, Comp. Geosci., *33* (8),1036–1051. doi: 10.1016/j.cageo.2006.11.003.

MATHWORKS (2002), *Matlab*, Version 6.5 Release 13, The MathWorks, Inc.

MEEHL, G.A., STOCKER, T.F., COLLINS, W.D., FRIEDLINGSTEIN, P., GAYE, A.T., GREGORY, J.M., KITOH, A., KNUTTI, R., MURPHY, J.M., NODA, A., RAPER, S.C.B., WATTERSON, I.G., WEAVER, A.J., and ZHAO, Z.-C. (2007), Global Climate Projections. In Climate Change 2007: The Physical Science Basis. Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L., (Eds.), Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, USA.

MEIER, M.F., and WAHR, J.M. (2002), *Sea level is rising: Do we know why*? Proc. Nat. Acad. Sci. *99(10)*, pp. 6524-6526.

NAGY, D., PAPP, G., and BENEDEK, J. (2000), *The gravitational potential and its derivatives for the prism*, J. Geodesy *74,* 552–560.

SOLOMON, S., QIN, D., MANNING, M., ALLEY, R.B., BERNTSEN, T., BINDOFF, N.L., CHEN, Z., CHIDTHAISONG, A., GREGORY, J.M., HEGERL, G.C., HEIMANN, M., HEWITSON, B., HOSKINS, B.J., JOOS, F., JOUZEL, J., KATTSOV, V.,

LOHMANN, U., MATSUNO, T., MOLINA, M., NICHOLLS, N., OVERPECK, J., RAGA, G., RAMASWAMY, V., REN, J., RUSTICUCCI, M., SOMERVILLE, R., STOCKER, T.F., WHETTON, P., WOOD, R.A., and WRATT, D. (2007), Technical Summary. In *Climate Change 2007: The Physical Science Basis*. Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L., (Eds.), Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, USA, pp. 19–91.

TORGE, W., *Geodäsie* (Walter de Gruyter GmbH & Co KG, Berlin 2003), 369 pages.

WEAVER, A.J., EBY, M., WIEBE, C.E., BITZ, C.M., DUFFY, P.B., EWEN, T.L., FANNING, A.F., HOLLAND, M.M., MACFADYEN, A., MATTHEWS, H.D., MEISSNER, K.J., SAENKO, O., SCHMITTNER, A., WANG, H., and YOSHIMORI, M. (2001), *The UVic Earth System Climate Model: Model description, climatology, and applications to past, present and future climates*, Atmosphere-Ocean *39* (4), 361–428.

WILLIS, J.K., Roemmich, D., and Cornuelle, B. (2004), *Interannual variability in upper-ocean heat content, temperature and thermosteric expansion on global scales*, J. Geophys. Res. *109*, C12036. doi: 10.1029/2003JC002260.

(Received January 23, 2007, revised February 26, 2008, accepted March 19, 2008)
Published Online First: July 11, 2008

Pure appl. geophys. 165 (2008) 1153–1165
0033–4553/08/061153–13
DOI 10.1007/s00024-004-0351-7

Pure and Applied Geophysics

A Fractal Interaction Model for Winding Paths through Complex Distributions: Application to Soil Drainage Networks

MIGUEL ÁNGEL MARTÍN,[1] and MIGUEL REYES[2]

Abstract—Water interacts with soil through pore channels putting mineral constituents and pollutants into solution. The irregularity of pore boundaries and the heterogeneity of distribution of soil minerals and contaminants are, among others, two factors influencing that interaction and, consequently, the leaching of chemicals and the dispersion of solute throughout the soil.

This paper deals with the interaction of irregular winding dragging paths through soil complex distributions. A mathematical modelling of the interplay between multifractal distributions of mineral/pollutants in soil and fractal pore networks is presented.

A Hölder path is used as a model of soil pore network and a multifractal measure as a model of soil complex distribution, obtaining a mathematical result which shows that the Hölder exponent of the path and the entropy dimension of the distribution may be used to quantify such interplay. Practical interpretation and potential applications of the above result in the context of soil are discussed. Since estimates of the value of both parameters can be obtained from field and laboratory data, hopefully this mathematical modelling might prove useful in the study of solute dispersion processes in soil.

Key words: Hölder curves, multifractal distributions, soil drainage networks.

1. Introduction

River basins distribute water stored in the soil by releasing it gradually into a complex network that involves a great disparity of length scales from the soil pore channels to the river basin boundaries. Along the way, water interacts with the basin, putting mineral constituent or soil pollutants into solution.

Percolation network theory and fractal models recently have been used for modelling the spreading of solute through porous media during saturated flow. ADLER (1985) considered dispersion in fractal capillary networks, REDNER *et al*. (1987) studied mechanical dispersion in a self-similar model of a porous medium, and MAZO (1998) studied different aspects of dispersion in fractal media. Fractal curves and networks appear as

[1] Departamento de Matemática Aplicada, Escuela Técnica Superior de Ingenieros Agrónomos, Universidad Politécnica de Madrid, 28040 Madrid, Spain. E-mail: miguelangel.martin@upm.es

[2] Departamento de Matemática Aplicada, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain. E-mail: mreyes@fi.upm.es

natural models for irregular geophysical boundaries, river basins or percolation channels (MANDELBROT, 1982; FEDER, 1988; RODRÍGUEZ-ITURBE and RINALDO, 1997).

On the other hand, distributions of nutrients and pollutants in soil demonstrate, as many other soil properties, high spatial and temporal variability. KRAVCHENKO *et al.* (1999) have shown that spatial distributions of soil phosphorus and potassium contents, organic matter contents, calcium and magnesium contents, and cation exchange capacity present highly heterogeneous patterns close to mathematical multifractal distributions. Similarly, mineral deposits have been shown to follow multifractal features (CHENG *et al.*, 1994; AGTERBERG *et al.*, 1996). These observations imply that the distributions are generally sparse, and denser and rarer regions follow certain scaling regularity. Such heterogeneity should have profound influence on leaching of metals and nutrients, dispersion of solutes throughout the soil, and other transport processes. Both, pore space geometry and the probability that particles or molecules have of being moved from the soil matrix to flow into the pore space, will affect solute transport (PERFECT and SUKOP, 2001). Thus, it is worthwhile to consider modelling the interplay between the complex geometry of pore networks and the heterogeneity of the mass distribution of soil components able to be transported through pore channels during saturated flow.

The objective of this work was to model such mass-geometry interplay using fractal dimension of branched transport pathways and the entropy dimension of spatial distributions of solute concentrations. In section 2 the model is developed and a mathematical result for the model is presented. A precise original mathematical proof of the theoretical result, supporting the value of this modelling, is given. In section 3 the practical interpretation and potential applications of the above result in the context of soil are discussed.

2. *A Fractal Interaction Model for Winding Dragging Paths through Soil Complex Distributions*

Mathematically speaking, a fractal network is a connected (possibly self-intersecting) curve or path of fractal (Hausdorff) dimension D. In an abstract setting, this concept directly corresponds with that of Hölder exponent of a continuous path. Since self-similar connected sets of points can be parameterized by means of continuous (Hölder) paths of the same Hausdorff dimension (REMES, 1998; MARTIN AND MATTILA, 2000) we shall model winding pore channels as the image of a Hölder map $f : A \longrightarrow \mathbb{R}^3$, $A \subset \mathbb{R}$, that is, a map verifying $|f(x) - f(y)| \leq c|x - y|^{\alpha}$ for all $x, y \in A$, with $0 < \alpha \leq 1$ and $c < \infty$. This model will allow us to obtain exact mathematical results that shall be interpreted later in a practical setting.

On the other hand the entropy dimension is a classical parameter used to quantify heterogeneity of mass distributions (RENYI, 1957) that may be estimated in real distributions by means of multifractal analysis of field data (see next section).

The above abstract modelling will allow us to give in this section a precise original proof of the following result which shows how the entropy dimension D_1 of the distribution and the Hölder exponent α of the path, play an important role in measuring the physical interplay between both structures.

Result: Let μ *be a multifractal measure being* D_1 *the entropy dimension. Suppose further that* $f : A \longrightarrow \mathbb{R}^n$ *is an* α*-Hölder map such that* $\gamma = f(\mathbb{R}) \subset S$ *and* $1/\alpha < D_1$. *Then* $\mu(\gamma) = 0$.

2.1. Preliminaries

The mathematical result is crucial for supporting the model. Next we present precise definitions and previous results needed.

Given a finite measure μ on $\mathbb{R}^n$ (or mass distribution), the local dimension (or local Hölder exponent) of μ at $x \in \mathbb{R}^n$ is given by (see e.g., FALCONER, 1997)

$$\dim_{loc} \mu(x) = \lim_{r \downarrow 0} \frac{\log \mu(B(x,r))}{\log r}$$

if this limit exists, where $B(x,r)$ denotes the closed ball $B(x,r) = \{y \in \mathbb{R}^n : |y - x| \leq r\}, x \in \mathbb{R}^n$ and $0 < r < \infty$.

For $0 \leq s \leq n$ the s-dimensional Hausdorff measure of a set $E \subset \mathbb{R}^n$ is

$$H^s(E) = \lim_{\delta \downarrow 0} \inf \left\{ \sum_{i=1}^{\infty} d(S_i)^s : E \subset \bigcup_{i=1}^{\infty} S_i, d(S_i) \leq \delta \right\}.$$

In particular, the Hausdorff measure H^n is a constant multiple of the Lebesgue measure $\mathcal{L}^n$.

If E is the support of the measure μ and $\mathcal{P} = \{A_i : i = 1, \ldots, n\}$ is a partition of E, the *Shannon entropy* of μ with respect to $\mathcal{P}$ is given by (SHANNON, 1948)

$$H_\mu(\mathcal{P}) = -\sum_{i=1}^{n} \mu(A_i) \log \mu(A_i)$$

If

$$H_\mu(\varepsilon) = \inf\{H_\mu(\mathcal{P}) : d(\mathcal{P}) \leq \varepsilon\}$$

being $d(\mathcal{P}) = \max_{1 \leq i \leq n}\{d(A_i)\}$, where d stands for the diameter, the *entropy dimension* of μ is defined by (RENYI, 1957)

$$D_1 = \lim_{\varepsilon \to 0} \frac{H_\mu(\varepsilon)}{-\log \varepsilon}$$

The *Hausdorff dimension of a set* $E \subset \mathbb{R}^n$ is defined by

$$\dim_H E = \inf\{s : H^s(E) = 0\} = \sup\{s : H^s(E) = \infty\}.$$

If $m \leq n$, $0 < \alpha \leq m$, and $A \subset \mathbb{R}^m$ we shall denote by $Lip_\alpha(A, \mathbb{R}^n)$ the set of Hölder continuous maps $f : A \longrightarrow \mathbb{R}^n$, that is

$$Lip_\alpha(A, \mathbb{R}^n) = \{f : A \longrightarrow \mathbb{R}^n : \exists L < \infty \text{ with } |f(x) - f(y)| \leq L|x - y|^\alpha, \ \forall x, y \in A\}.$$

The number

$$L = \sup\left\{\frac{|f(x) - f(y)|}{|x - y|^\alpha} : x, y \in A, x \neq y\right\}$$

is called the *Hölder constant* of *f*.

The next theorem plays an important role in the proof of our results (see MATTILA, 1995).

2.2. Theorem (Besicovitch's Covering Theorem)

There are integers P(n) and Q(n) depending only on n with the following properties. Let A be a bounded subset of $\mathbb{R}^n$, *and let* $\mathcal{B}$ *be a family of closed balls such that each point of A is the center of some ball of* $\mathcal{B}$.

a) *There is a finite or countable collection of balls* $\{B_i\} \subset \mathcal{B}$ *such that they cover A and every point of* $\mathbb{R}^n$ *belongs to at most P(n) balls* B_i, *that is*,

$$\chi_A \leq \sum_i \chi_{B_i} \leq P(n)$$

where χ_A *denotes the characteristic function of A.*

b) *There are families* $\mathcal{B}_1, \ldots, \mathcal{B}_{Q(n)} \subset \mathcal{B}$ *covering A such that each* $\mathcal{B}_i$ *is disjoint, that is*,

$$A \subset \bigcup_{i=1}^{Q(n)} \bigcup_{B \in \mathcal{B}_i} B \quad \text{and} \quad B \cap B' = \emptyset \quad \text{for} \quad B, B' \in \mathcal{B}_i \quad \text{with} \quad B \neq B'$$

A Hölder map $f : A \longrightarrow \mathbb{R}^3$, $A \subset \mathbb{R}$, gives a parameterization of a fractal path (MATTILA, 1995). If the image *f*(*A*) is embedded in the support *S* of a mass distribution, one natural problem is to relate the heterogeneity of the mass distribution, being this measured via the entropy dimension, with the Hölder exponent of the map, in order to create the possibility that the image *f* (*A*) can catch a positive amount of mass. The next theorem deals with this problem. It is presented in a general form for distributions in $\mathbb{R}^n$.

2.3. Theorem

Let μ *be a measure supported on* $E \subset \mathbb{R}^n$, *with entropy dimension* D_1, *and* $0 < s < D_1$. *Then, for any (m/s)-Hölder map* $f : A \longrightarrow \mathbb{R}^n$, $A \subset \mathbb{R}^m$, *we have*

$$\mu(E \cap f(A)) = 0.$$

Proof. We may assume that $d(E) < 1$. Since the entropy dimension is D_1, then (YOUNG, 1982):

$$\lim_{r\to 0} \frac{\log \mu(B(x,r))}{\log r} = D_1 \quad \text{at} \quad \mu - \text{almost all points} \quad x \in E.$$

If $s < D_1$ is easy to show that

$$\lim_{r\to 0} \frac{\mu(B(x,r))}{r^s} = 0 \quad \text{at} \quad \mu - \text{almost all points} \quad x \in E.$$

Presume now, contrary to the assertion, that for some $s < D_1$ there exists an (m/s)-Hölder map $f : A \longrightarrow \mathbb{R}^n$, $A \subset \mathbb{R}^m$, with $\mu(E \cap f(A)) > 0$. Take $\delta > 0$ such that $\mu(E \cap f(A)) > \delta$. By Egorov's theorem, there is $E_\delta \subset E$ such that $\mu(E_\delta) > \mu(E) - \delta$ and

$$\lim_{r\to 0} \frac{\mu(B(x,r))}{r^s} = 0 \quad \text{uniformly on} \quad E_\delta.$$

Moreover, since $\mu(E \cap f(A)) > \delta$, then $\mu(E_\delta \cap f(A)) > 0$.

Since Hölder maps can be extended (see VI.2.2 in STEIN, 1970) it may be surmesed that A is open and that $\mathcal{L}^m(A) < \infty$.

Since μ-almost all points of $E_\delta \cap f(A)$ are μ-density points (see, for example, 2.14 in MATTILA, 1995), then for μ-almost all $x \in E_\delta \cap f(A)$ one has

$$\lim_{r\to 0} \frac{\mu(E_\delta \cap f(A) \cap B(x,r))}{\mu(B(x,r))} = 1$$

and then

$$\lim_{r\to 0} \frac{\mu(E_\delta \cap f(A) \cap B(x,r))}{r^s} = 0$$

Let $\varepsilon > 0$ arbitrary. Then there is $R > 0$ such that

$$\frac{\mu(E_\delta \cap f(A) \cap B(x,r))}{r^s} < \varepsilon \tag{1}$$

for $r \leq R$ and for μ-almost all $x \in E_\delta \cap f(A)$.

Let consider now the covering of $E_\delta \cap f(A)$ formed by balls $B(x,R)$, $x \in E_\delta$. Applying Besicovith covering theorem 2.2, we can get a sequence of balls $\{B_k\}$ verifying (1) such that

$$E_\delta \cap f(A) \subset \bigcup_k B_k$$

Moreover, that sequence may be grouped in a finite number of families $\mathcal{B}_1, \ldots, \mathcal{B}_{Q(n)}$ being $Q(n)$ a constant depending only on n, such that $B \cap B' = \emptyset$ for $B, B' \in \mathcal{B}_j$ with $B \neq B'$.

It follows that there is at least one of these families, say $\mathcal{B}_j$, such that

$$\sum_{B_i \in \mathcal{B}_j} \mu(B_i \cap E_\delta \cap f(A)) \geq \frac{1}{Q(n)} \mu(E_\delta \cap f(A))$$

Since the balls B_i all have the same radius, the family of balls $\mathcal{B}_j$ is finite. Say $\mathcal{B}_j = \{B_1, \ldots, B_N\}$. Thus one has

$$N\ \varepsilon\ R^s \geq \sum_{i=1}^{N} \mu(B_i \cap E_\delta \cap f(A)) \geq \frac{1}{Q(n)} \mu(E_\delta \cap f(A))$$

and thus

$$N \geq \frac{\mu(E_\delta \cap f(A))}{\varepsilon Q(n)} R^{-s}$$

Let $B_i = B(x_i, R_\varepsilon)$ with $x_i \in E_\delta \cap f(A)$, $A_i = f^{-1}(B_i)$ and $y_i \in A_i$ with $f(y_i) = x_i$. The Hölder condition implies that

$$B'_i = B\left(y_i, L^{-s/m} R^{s/m}\right) \subset A_i \subset A \quad \text{and} \quad \mathcal{L}^m(B'_i) \geq c_m R^s.$$

Moreover, since the balls $\{B_1, \ldots, B_N\}$ are disjoints, the balls $\{B'_1, \ldots, B'_N\}$ are also disjoints. Then

$$\mathcal{L}^m(A) \geq \sum_{i=1}^{N} \mathcal{L}^m(B'_i) \geq \frac{\mu(E_\delta \cap f(A))}{\varepsilon Q(n)} R^{-s} c_m R^s = \frac{\mu(E_\delta \cap f(A)) c_m}{\varepsilon Q(n)}$$

Since $\mu(E_\delta \cap f(A)) > 0$, $c_m > 0$ and $\varepsilon > 0$ is arbitrary, then $\mathcal{L}^m(A) = \infty$ which is a contradiction, and the statement follows. □

2.4. Remark

Notice that rectifiability properties are studied above through coverings formed by balls of equal radius, due to the use of Besicovitch covering theorems instead of Vitali's type that render covering by balls of different sizes. For distributions coming from computer simulation of dynamical systems or else experimental distributions, this seems more convenient: one may not only be interested in limiting properties as rectifiability but also in scaling properties of coverings of controlled size, as obtained in the proof of theorem 2.3.

The entropy dimension thus appears as a degree of accessibility to the mass through continuous paths, giving a measure of the tortuosity needed to catch an important amount of mass.

Invariant measures of dynamical systems produce typical examples of multifractal measures (Pesin, 1996). In the important case of self-similar measures, the result above takes a specific parameterized formulation. Namely, if $\{f_1, \ldots, f_N\}$ are contractions in $\mathbb{R}^n$

and $\{p_1,\ldots, p_N\}$ are positive numbers such that $\sum_{i=1}^{N} p_i = 1$, there is a unique measure verifying

$$\mu = \sum_{i=1}^{N} p_i \mu \circ f_i^{-1},$$

which is called the invariant measure associated to the *iterated function system* $\{f_1,\ldots,f_N; p_1,\ldots, p_N\}$. In the case that the $\{f_1,\ldots, f_N\}$ are similarities one has the following result.

2.5. Corollary

Let $\{f_1,\ldots,f_N\}$ *be similarities in* $\mathbb{R}^n$ *with contraction ratios* $\{r_1,\ldots,r_N\}, \{p_1,\ldots,p_N\}$ *positive numbers such that* $\sum_{i=1}^{N} p_i = 1$, *and let* μ *be the invariant measure with respect to the iterated function system* $\{f_1,\ldots,f_N; p_1,\ldots,p_N\}$. *Suppose that* $\mu\big(f_i(E) \cap f_j(E)\big) = 0$ *for* $i \neq j$ *being* E *the support of* μ. *Then if*

$$0 < s < \frac{\sum_{i=1}^{N} p_i \log p_i}{\sum_{i=1}^{N} p_i \log r_i}$$

for any (m/s)*-Hölder map* $f : A \longrightarrow \mathbb{R}^n$, $A \subset \mathbb{R}^m$, *we have* $\mu(E \cap f(A)) = 0$.

Proof. It is a direct consequence of the fact that the entropy dimension of self-similar measures is given by the formula (Deliu *et al*., 1991):

$$D_1 = \frac{\sum_{i=1}^{N} p_i \log p_i}{\sum_{i=1}^{N} p_i \log r_i}$$

□

3. Applications to Soil Drainage Networks

Field data corresponding to soil properties can be collected in one, two or three spatial dimensions. The characteristics (i.e., shape, size and connectivity) of pore networks are often studied by two-dimensional image analysis of thin sections. Similarly, information on the spatial variability of soil properties is usually collected at different sites (points) located along a transect or over a given area. Thus, fractal modelling of the boundaries of pore channels and capillaries can be made by means of fractal curves and networks. Analogously, mineral or contaminant concentrations in soil may be represented by means of a mass distribution. We develop in this section a mass-geometric fractal modelling of the interplay between pore space geometry and the mass distribution of nutrients and contaminants.

3.1. Fractal Pore Channel Networks

Soil is formed by an intricate arrangement of solid particles and voids (pores) with connecting pore channels through which fluid flow and solute transport take place. A number of different approaches have been used to model this situation (JURY and FLÜER, 1992), including those which apply fractal and percolation models (SAHIMI, 1993; ADLER, 1985; among others). The fractal nature of pore boundaries within a range of scales has been demonstrated. The concept of tortuosity applied to pore channels, widely used in soil sciences, has a precise meaning in terms of the scaling behavior of the length of pores. Although a wide variety of models has been used to describe pore geometry (PERFECT and SUKOP, 2001) an ideal pore channel may be modelled by a fractal curve.

In order to quantify pore channel tortuosity the boundary fractal dimension D is used which is defined by means of the scaling equation (KAMPICHCHLER and HAUSER, 1993; ANDERSON *et al*., 1998; PACHEPSKY *et al*., 1996):

$$L(\varepsilon) \approx L_1 \varepsilon^{1-D}$$

where $L(\varepsilon)$ is the measured length using a yardstick of normalized length ε, and L_1 is the measured length when ε is equal to unity.

Different values for the fractal dimension of pore boundaries ranging from 1.06 to 1.51 have been found. Also the effect of management practices on such values has been studied (PACHEPSKY *et al*., 1996).

3.2. Parameterizing Heterogeneity of Soil Distributions

The distribution of soil mineral components and pollutants also show a high spatial variability. Since water interacts with soil mineral constituents and pollutants through soil pore channels, the spatial variability of their concentrations is a factor that should be taken into account in the study of solute dispersion processes. Thus, the geometry of the distribution of some soil minerals or contaminants is a crucial feature to determine the accessibility of pore channels to disperse soil componentes. For studying those distributions one may consider a *measure* or *distribution* that assigns to every region E the quantity $\mu(E)$ of a certain component located in that region. Typically $\mu(E)$ depends on the location of that region of the medium and varies widely with respect to the volume of E, having the main features of multifractal measures. This implies that, being sparse within the solid matrix, there exists denser and rarer regions following certain scaling regularity. In order to characterize the complexity of the spatial distribution, multifractal analysis may be used to estimate the Renyi spectrum of dimensions which include the entropy dimension as a significant dimension (see EVERSTZ and MANDELBROT, 1992).

Let S be the support of a distribution μ (for methodological reasons we use a two-dimensional model here). Let $\mathcal{P} = \{R_i\}_{i=1}^{N}$ be a collection of squares of side length ε (see Fig. 1) that represent a partition of S.

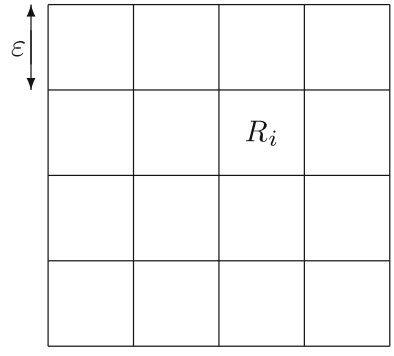


Figure 1

The Renyi dimensions are defined as

$$D_q = \frac{1}{q-1}\lim_{\varepsilon\to 0}\frac{\log\sum_{i=1}^{N(\varepsilon)}\mu_i(\varepsilon)^q}{\log\varepsilon},$$

for $q \neq 1$, being D_1 the entropy dimension computed by

$$D_1 = \lim_{\varepsilon\to 0}\frac{\sum_{i=1}^{N(\varepsilon)}\mu_i(\varepsilon)\log\mu_i(\varepsilon)}{\log\varepsilon}.$$

The dimension D_0 is called the capacity dimension which agrees with the fractal dimension of the support S.

When D_q is a decreasing function of q, μ is called a *multifractal distribution*.

3.3. Application to Soil Drainage

The theoretical result of section 2 may be interpreted in the practical context of soil. It suggests to use the exponent $D_1 - D$ as an indicator that might reflect the likelihood of leaching or dispersion of soil chemicals and minerals, based solely on the physical interplay between the pore boundary and chemical molecules. The greater the index is the higher is the probability of chemical molecules to be dispersed into the water pore channel. This probability would diminish when the difference $D_1 - D$ approaches zero and increases when it becomes negative. The greater the entropy dimension is, the more tortuosity for the channel is needed, and thus this parameter may be used as a measure of the risk of mineral loss by illuviation.

In the case of soil contaminants it would measure the risk of exporting pollution to the surrounding areas. Particle size soil distributions have been shown to obey fractal scaling laws (TURCOTTE, 1986; TYLER and WHEATCRAFT, 1990) and the power-scaling exponent has been related with tortuosity (TYLER and WHEATCRAFT, 1989).

Thus the knowledge of such exponent or other entropy-like quantities (MARTÍN *et al.*, 2001) characterizing texture, together with the entropy dimension of certain distributions of soil mineral, provide valuable information that may be used for the diagnosis of soils and eventually may be of help in implementing adequate policies. In this sense, the potential application might be wide and have a real value in a practical setting. It is well known that parameters and meaningful indexes, such as those above, are strongly demanded by soil scientists (DORAN and PARKIN, 1994).

3.4. A Case Study

Soil samples corresponding to an agricultural field (vineyard) located in Central Spain have been collected. A total number of 256 sampling points in an square lattice was considered, being the distance between two neighboring points equal to 20 meters. At any point of the lattice a sample at 25 cm. depth was taken and potassium, phosphorous and organic matter contents were obtained by laboratory standard techniques. This produces data sets $\{\mu_i : i = 1, 2,..., 256\}$ corresponding to the respective potassium/phosphorous and organic matter contents.

The probability measure or mass distribution μ is constructed, assigning to any subsquare $R \subset S$ a measure or mass

$$\mu(R) = \frac{\sum_{x_i \in R} \mu_i}{\sum_{x_i} \mu_i}$$

At every $\mu_i(\varepsilon)$ for $i = 1, 2,..., N(\varepsilon)$ is computed.

The multifractal dimensions D_q are estimated by a least-square fitting of the corresponding scalings endowed in equation (1), for ε ranging from $\varepsilon = 16$ to $\varepsilon = 1$, and q ranging from $q = -10$ to $q = 10$ with a lag of 0.5. The multifractal dimensions D_q are plotted against q in Figures 2 and 3.

Coefficients of determination (R^2) in those fittings and D_q values range from 0.993 to 0.999 for q values ranging from $q = -10$ to $q = 10$.

The estimated values of the entropy dimensions of potassium and phosphorous distributions were 1.973 and 1.964 respectively which means that both distributions have followed quite similar heterogeneity patterns. It would mean that both minerals have a very similar probability to be in contact with soil water during saturated flow.

Taking into account the increase of the pore boundary dimension value after tillage and other management practices (PACHEPSKY *et al.*, 1996) the exponent $D_1 - D$ would be affected and consequently the probability of leaching of different minerals by the effect of watering or rain events. It follows that the use of parameters proposed here might be useful to assess the planning of management practices in agricultural fields.

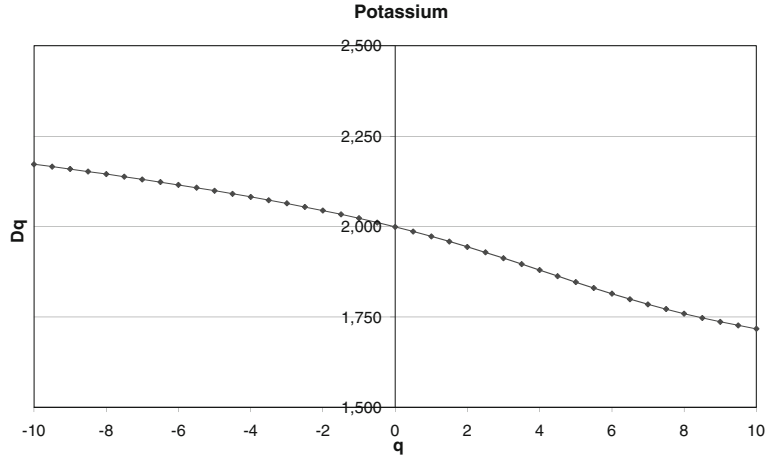


Figure 2

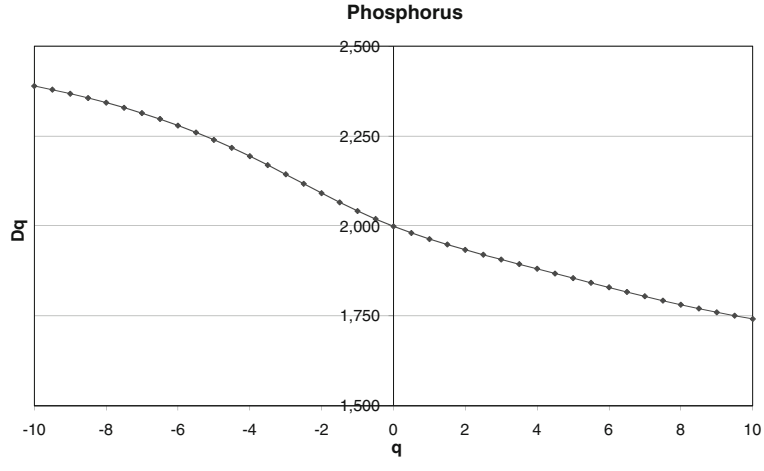


Figure 3

4. Conclusions

Physical and chemical heterogeneities coexist together and interplay in soil. Under a modelling of this interplay, the boundary of pore space may be characterized by the fractal dimension D and the heterogeneity of soil minerals or contaminants chemical may be characterized by the entropy dimension D_1. A mathematical result is precisely derived and interpreted in a practical context. If $D << D_1$ the pore network has only a small

probability of catching the chemical mass spread in a heterogeneous multifractal manner. This probability would diminish when the difference $D_1 - D$ approaches zero and increases when it becomes negative. Although this is a simple and schematic modelling of a considerably more complex situation, this result relates to factors that influence solute dispersion and appear unconnected in former studies. Since estimates of the value of both parameters can be obtained from laboratory and field data, hopefully this mathematical modelling might be useful in future studies

Acknowledgments

This work has partially been supported by Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I) (Spain), Ref.: AGL2007-62648, and by DGUI (Comunidad de Madrid) - UPM (Technical University of Madrid), Ref.: M0700204135.

References

Adler, P.M. (1985), *Transport processes in fractals. III. Taylor dispersion in two examples of fractal capillary networks*, Int. J. Multiphase Flow *11*, 241–254.

Agterberg, F.P., Cheng, O., Brown, A., and Good, D. (1996), *Multifractal modelling of fractures in the Lac Du Bonnet batholith, Manitoba*, Comput. Geosci. *22*, 497–507.

Anderson, A.N., McBratney, A.D., and Crawford, J.W. (1998), *Application of fractals to soil studies*, Advances in Agron. *63*, 1-76.

Cheng, O., Agterberg, F.P., and Ballantyne, S.B. (1994), *The separation of geochemical anomalies from background by fractal models*, J. Geochem. Explor. *51*, 109–130.

Deliu, A., Geronimo, J.S., Shonkwiler, R., and Hardin, D. (1991), *Dimensions associated with recurrent self-similar sets*, Math. Proc. Cambridge Philos. Soc. *110*, 327–336.

Doran, J.W. and Parkin, T.B., *Defining and assessing soil quality*. In *Defining Soil Quality for a Sustainable Environment* (eds. Doran, J.W., Coleman, D.C., Bezdicek, D.F., and Stewart, B.A.). (SSSA Special Publication number 35, Madison, Wisconsin, USA. Soil Science Society of America 1994) pp. 3–22.

Evertsz, C.F.G. and Mandelbrot, B.B., *Multifractal measures*. In *Chaos and Fractals* (eds. Peitgen, H.O., Jürgens, H. and Saupe, D.) (Springer Verlag, New York 1992) pp. 921–953.

Falconer, K.J., *Techniques in Fractal Geometry* (John Wiley & Sons, Chichester 1997).

Feder, J., *Fractals* (Plenum Press, New York 1988).

Jury, W.A. and Flühler, H. (1992), *Transport of chemicals through soil: Mechanisms, models and field applications*, Adv. Agron. *47*, 141–201.

Kampichchler, C. and Hauser, M. (1993), *Roughness of soil pore surface and its effect on available habitat space of microarthropods*, Geoderma *56*, 223–232.

Kravchenko, A.N., Boast, C.W., and Bullock, D.G. (1999), *Multifractal analysis of soil spatial variability*, Agron. J. *91*, 1033–1041.

Mandelbrot, B.B., *Fractal Geometry of Nature* (Freeman, New York 1982).

Martín, M.A. and Mattila, P. (2000), *On the parameterization of self-similar and other fractal sets*, Proc. Am. Math. Soc. *128*, 2641–2648.

Martín, M.A., Rey, J.M. and Taguas, F.J. (2001), *An entropy-based parameterization of soil texture via fractal modelling of particle-size distribution*, Proc. R. Soc. London Ser. A *457*, 937–948.

Mattila, P., *Geometry of Sets and Measures in Euclidean Spaces* (Cambridge University Press, Cambridge 1995).

MAZO, R.M. (1998), *Taylor dispersion on a fractal*, Acta Phys.Polon. B *29*, 1539–1549.

PACHEPSKY, Y., YAKOVCHENKO, V., RABENHOST, M.C., POOLEY, C. and SIKORA, L.J. (1996), *Fractal parameters of pore surface as derived from micromorphological data: effect of long-term management practices*, Geoderma *74*, 305–319.

PERFECT, E. and SUKOP, M.C., *Models relating solute dispersion to pore space geometry in satured media: A review*. In *Physical and Chemical Processes of Water and Solute Transport / Retention in Soil* (eds. Selim, H.M. and Sparks, D.L.). (Special Publ. 56, Soil Sci. Soc. Am., Madison WI 2001) pp. 77–146.

PESIN, Y., *Dimension Theory in Dynamical Systems* (Chicago University Press, Chicago 1996).

REDNER, S., KOPLIK, J. and WILKINSON, D. (1987), *Hydrodynamic dispersion in a self-similar geometry*, J. Phys. A *20*, 1543–1555.

REMES, M. (1998), *Hölder parameterization of self-similar sets*, Ann. Acad. Sci. Fenn. Ser A I Math, Dissertations *112*.

RENYI, A., *Dimension, entropy and information*. In *Statistical Decision Functions and Random Processes* (Trans. 2*nd* Prague Conf. on Information Theory, 1957) pp. 545–556.

RODRÍGUEZ-ITURBE, I. and RINALDO, A., *Fractal River Basins, Chance and Self-organization* (Cambridge University Press, Cambridge 1997).

SAHIMI, M. (1993), *Flow phenomena in rocks: From continuum models to fractals, percolation, cellular automata, and simulated annealing*, Modern Physics *65*, 1393–1534.

SHANNON, C.E. (1948), *A mathematical theory of communication I*, Bell Syst. Tech. J. *27*, 379–423.

STEIN, E.M., *Singular Integrals and Differentiability Properties of Functions* (Princeton University Press, Princeton 1970).

TURCOTTE, D.L. (1986), *Fractals and fragmentation*, J. Geophys. Res. *91*, 1921–1926.

TYLER, S.W. and WHEATCRAFT, S.W. (1989), *Application of fractal mathematics to soil water retention estimation*, Soil Sci. Soc. Am. J. *53*, 987–996.

TYLER, S.W. and WHEATCRAFT, S.W. (1990), *Fractal processes in soil water retention*, Water Resour. Res. *26*, 1047–1054.

YOUNG (1982), *Dimension, entropy and Lyapunov exponents*, Ergodic Theory Dynamical Systems *2*, 109–124.

(Received April 9, 2007, revised March 27, 2008, accepted March 27, 2008)

Pure appl. geophys. 165 (2008) 1167–1193
0033–4553/08/061167–27
DOI 10.1007/s00024-008-0352-6

Pure and Applied Geophysics

The Coherent Pixels Technique (CPT): An Advanced DInSAR Technique for Nonlinear Deformation Monitoring

Pablo Blanco-Sánchez,[1] Jordi J. Mallorquí,[1] Sergi Duque,[1] and Daniel Monells[1]

Abstract—This paper shows the potential applicability of orbital Synthetic Aperture Radar (SAR) Differential Interferometry (DInSAR) with multiple images for terrain deformation episodes monitoring. This paper is focused on the *Coherent Pixels Technique* (CPT) developed at the Remote Sensing Laboratory (RSLab) of the Universitat Politecnica de Catalunya (UPC). CPT is able to extract from a stack of differential interferograms the deformation evolution over vast areas during wide spans of time. The former is achieved thanks to the coverage provided by current SAR satellites, like ESA's ERS or ENVISAT, while the latter due to the large archive of images acquired since 1992. An interferogram is formed by the complex product of two SAR images (one complex conjugate) and its phase contains information relative to topography, terrain deformation and atmospheric conditions among others. The goal of differential interferometric processing is to retrieve and separate the different contributions. The processing scheme is composed of three main steps: firstly, the generation of the best interferogram set among all the available images of the zone under study; secondly, the selection of the pixels with reliable phase within the employed interferograms and, thirdly, their phase analysis to calculate, as the main result, their deformation time series within the observation period. In this paper, the *Coherent Pixels Technique* (CPT) is presented in detail as well as the result of its application in different scenarios. Results reveal its practical utility for detecting and reproducing deformation episodes, providing a valuable tool to the scientific community for the understanding of considerable geological process and to monitor the impact of underground human activity.

Key words: Orbital SAR, differential interferometry, deformation monitoring.

1. Introduction

A *Synthetic Aperture Radar* (*SAR*) is a coherent imaging sensor able to acquire high resolution images from orbital platforms independently of the weather or sunlight conditions (Hanssen, 2001). All its capabilities have made this technique one of the fundamentals for Earth observation over oceans as well as over land. The images obtained from a SAR are complex. The amplitude depends on the scene reflectivity while its phase, among others, is proportional to the two-way distance from satellite to ground and therefore to the geometry of the scene. The combination of two SAR images of the same scene acquired from different orbits, i.e., incidence angles, produces an

[1] Remote Sensing Laboratory (RSLab), Universidad Politecnica de Catalunya (UPC), Campus Nord UPC, Barcelona. E-mail: pblanco@tsc.upc.edu; mallorqui@tsc.upc.edu; sergi.duque@tsc.upc.edu

interferogram. The *interferogram* is obtained by multiplying one image by the complex conjugate of the other and contains, on a pixel by pixel basis, the phase difference between the two acquisitions (MASSONET *et al.*, 1993). There is the agreement of calling one of the images as *master* and the other as *slave*. This phase difference can be exploited in combination with the orbital information for each acquisition to derive a *Digital Elevation Model* (*DEM*) of the scene. This interferometric processing of SAR data (*InSAR*) was used for instance by NASA/DLR/ASI for the *Shuttle Radar Topographic Mission* (*SRTM*) to obtain elevation data on a near-global scale for generating the most complete high-resolution digital topographic database of the Earth. SRTM consisted of a specially modified radar system that flew onboard the Space Shuttle Endeavour during an 11-day mission during February 2000 (NASA).

The next step in SAR interferometry has been the detection of Earth surface movements with *Differential Interferometry* (*DInSAR*) that has shown excellent results in the last years of research. Initial single interferogram DInSAR techniques (MASSONET, 1993; PELTZER, 1995) have evolved to multi-image techniques which are able to retrieve the deformation movement of the studied areas with, at least theoretically, millimetric precision (BERARDINO *et al.*, 2002; FERRETTI *et al.*, 2000; MORA *et al.*, 2003; LANARI *et al.*, 2004). The application of such techniques has extended to many forms of surface deformation in seismology, volcanology, anthropogenic subsidence or uplift and glacier monitoring. The latest achievements of these techniques consist in merging data from different sensors, as ERS and ENVISAT data (MONTI-GUARNIERI *et al.*, 2000; BLANCO *et al.*, 2005). DInSAR techniques present three immediate advantages compared to other classical methods employed to measure deformation episodes, such as the *Differential Global Positioning System* (*DGPS*) or other instrumental methods. First, they provide at low cost wide coverage of the studied area in opposition to the discrete point measurements supplied by the instrumental techniques, which in general present a benchmark density and extension lower to the DInSAR techniques. For instance, a single ERS or ENVISAT image covers an area of 100 km by 100 km. Secondly, the orbital sensors have an almost monthly revisit time, which helps to perform continuous monitoring of the selected location. Finally, there is an archive of images since 1992 that allows to study, at least in Europe, almost any place since that date. Nevertheless, both techniques should be seen as complementary rather than opposite.

The basic idea behind these techniques is to observe the area along time while acquiring SAR images on a regular basis. These images are then combined to generate a set of differential interferograms. A differential interferogram is just a regular interferogram whose topographic component has been removed. The topographic contribution is calculated using the orbital information of the pair and an external DEM. In an ideal case, the differential interferogram should contain only the deformation during the times of acquisition of the two images, however in practice there are other terms that can hide the desired information. The goal of the different processing techniques is to accurately isolate the deformation term from the rest in the set of interferograms. The

interferometric phase of a single pixel, ϕ, can be decomposed in several terms as shown in the following expression (HANSSEN, 2001, MORA *et al*., 2003),

$$\phi^k(T, B_n, \Delta f_{dc}) = \frac{4\pi}{\lambda_k} \cdot (v \cdot T + \beta) + \frac{4\pi}{\lambda_k} \frac{B_n}{r_0 \sin\theta} \cdot \varepsilon - \frac{2\pi}{v_s} \Delta f_{dc} \cdot az + \eta_{atm} + n, \qquad (1)$$

where k stands for the sensor type (for instance, ERS or ENVISAT), λ_k is the wavelength, T and $B_n(T)$ are respectively the temporal and perpendicular baselines (i.e., temporal gap between acquisitions and spatial separation of the two orbits) of the *ith-interferogram*, v the linear velocity of deformation, β the nonlinear deformation term, r_0 the sensor-target distance (*range*), θ is the local incidence angle, ε the DEM error, v_s is the sensor velocity, Δf_{dc} is the Doppler centroid difference between the two images of the interferogram, *az* the point scatter azimuth position with respect to the center of its pixel, η_{atm} accounts for the atmospheric phase artifacts and n the decorrelation noise.

The first term is the contribution to the interferometric phase of the deformation experienced by the pixel during the time interval of the pair. It has been decomposed in two terms, linear and nonlinear, because this is the model CPT uses to retrieve the deformation from the available data. The second term stands for the phase due to the non-compensated topography when generating the differential interferogram. Any inaccuracy of the DEM leads to a residual topographic phase (called DEM error). It depends on the spatial baseline and the larger the separation between the orbits the more sensitive is the phase to the topography. The third term, the Doppler centroid phase term, is only considered when working with *Permanent Scatterers* (FERRETTI *et al*., 2001), this is point-like targets. When working with a coherence approach, with averaged (multi-looked) interferograms, only the common part of the azimuth spectrum of the images is used and consequently the filtering makes both images to share the same Doppler centroid. Under a coherence approach only interferograms with low Δf_{dc} are useful. The differences between both approaches are analyzed in section 3. The fourth term is known as *atmospheric phase screen* and it is caused by the different atmospheric propagation conditions in both images, i.e., changes in the wave speed, causing something similar to a ghost topography. Finally, the last term includes all sources of decorrelation and has to be regarded as a degradation of the phase quality. Decorrelation can be defined as any noise caused by error sources that have a small correlation length (few pixels). Among them, the most important are the following. The *baseline* or *geometric decorrelation* is caused by the difference in the incidence angles between the two acquisitions. It is inherent to interferometry and worsens with the spatial baseline length. Baselines larger than the so-called *critical one* lead to useless interferograms. The *temporal decorrelation* is caused by physical changes in the terrain that affect the scattering characteristics of the surface. It worsens with the temporal baseline, as terrain is more likely to change as time passes. The *volume decorrelation* is caused by the penetration of the radar wave in scattering media. It justifies why it is difficult to obtain good interferograms over forested areas even with short temporal baselines. The *Doppler centroid decorrelation* is caused by the differences of the Doppler centroids between the two acquisitions. The *thermal* or *system*

noise is caused by the noise that any microwaves hardware adds to the measured signal. Each source has assigned a *coherence* or *correlation*, which ranges from 0 (total decorrelation or pure noise) to 1 (complete correlation or noiseless case). The total correlation or coherence is simply the product of all terms (HANSSEN, 2001).

All DInSAR techniques share the same principles. In particular the *Coherent Pixels Technique* (*CPT*) (MORA *et al*., 2003; BLANCO *et al*., 2006), will be presented as well as a variety of results achieved under different scenarios and environmental conditions. CPT can be divided in four main blocks. The first two refer to the *interferogram set generation* and the *quality pixels selection*, this is how to properly select the data to be processed from all the available. The last two refer to the way CPT extracts the useful information, i.e., the deformation time series, from the available data, i.e., the differential interferograms stack. Following a divide and conquer strategy, the linear and nonlinear terms of the differential phase are calculated into two separate (but complementary) blocks.

The paper is structured as follows. In section 2 different approaches for interferogram set selection are presented. The different pixel selection criteria are presented in section 3, making a clear separation between amplitude and coherence-based approaches. In section 4 the linear block is presented and in section 5 the nonlinear, which obtain the linear term of deformation and the nonlinear one respectively. Section 6 presents some results of deformation time-series obtained with CPT, processing datasets from different sites. Finally, conclusions are outlined in section 7.

2. *Interferogram Set Selection Method*

The first option to create the set of interferograms would be to perform all the possible combinations between the available images. Nevertheless, the huge number of resultant combinations prevents us from doing so. In addition, some of the combinations may lead to useless interferograms when working with coherence-based techniques due to excessive spatial or temporal baselines or incompatible Doppler centroid frequencies of the pair. In order to perform a more optimized selection, CPT uses a 3D Delaunay triangulation (LEE *et al*., 1980) of the available images in the space defined by the spatial baseline, the temporal baseline and the Doppler centroid frequency. This triangulation relates all images with the minimum spatial and temporal baselines and Doppler difference which maximizes the phase quality of the generated interferograms. As the triangles cannot overlap each other, the total number of interferograms is clearly reduced. Furthermore, the triangulation helps to achieve a uniform distribution of spatial and temporal baselines that maximizes the quality of the results.

Once the triangulation has been performed it is advisable to remove all those interferograms with spatial or temporal baselines or Doppler frequencies over a maximum value in order to preserve their phase quality. For instance, in most of the results included in this paper the maximum values have been of the order of $B_n \sim 300$ m,

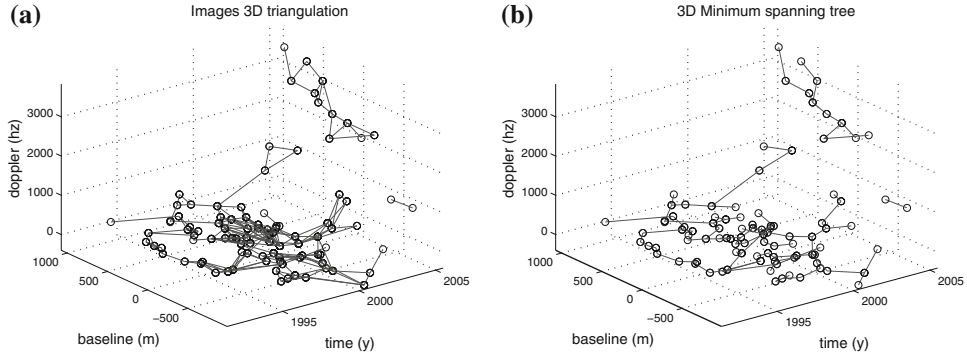


Figure 1
(a) 3D representation of the temporal baseline, spatial baseline and Doppler frequency distributions of the 268 interferograms by means of the Delaunay's triangulation. Lines connect the images which constitute the interferograms and delineate image subsets and (b) 3D representation of the temporal baseline, spatial baseline and Doppler frequency distributions of the minimum spanning tree of the initial 3D triangulation.

$T \sim 3$ years and $\Delta f_{dc} \sim 500$ Hz. As we will see, this is especially important when working with coherence-based pixel selection criteria, while for amplitude-based these restrictions can be relaxed or even skipped. The 3D triangulation for the Gardanne (South of France) test site images is depicted in Figure 1(a). From this triangulation, 268 interferograms are created (255 ERS-ERS and 13 ENVISAT-ENVISAT) out of 93 images (85 ERS and 8 ENVISAT). Four image subsets are obtained. Subsets are groups of interferograms having no common images with the rest. Two of the subsets are due to the different carrier frequencies of ERS and ENVISAT, as useful cross-interferograms can only be done under very special conditions (GATELLI *et al*., 1994; MONTI-GUARNIERI *et al*. 2000), while the other two are due to the commented restrictions.

Even with the 3D triangulation the number of interferograms is usually much larger than the minimum number required, this is $M - 1$ linearly independent being M the number of available images. Although redundant information helps to improve the results, an excess of interferograms leads to larger computational and disk storage requirements. For this reason an optimization over the 3D triangulation can be performed. The goal is to obtain $M - 1$ linearly independent interferograms by selecting the best combination in terms of phase quality. The proposed selection is based on finding the *Minimum Spanning Tree*, *MST* (AHUJA *et al*., 1993), of the 3D triangulation where the *interferometric phase coherence* will be used to compute a cost function to minimize. The triangulation is treated as a graph where the nodes are the images and the links the interferograms. A *spanning tree* of that graph is a subgraph which is a tree and connects all the images together (a single graph can have many different spanning trees). Each link has assigned a weight that represents how unfavorable it is, and any spanning tree has a cost function obtained from the sum of all weights of the links in it. A *Minimum Spanning Tree* or *Minimum Weight Spanning tree*

is then a spanning tree with a cost function less than or equal to the cost function of every other spanning tree. In order to implement this, the *interferometric phase coherence* is theoretically calculated from the mathematical expressions of the different correlation factors affected by the spatial and temporal baselines and the Doppler centroid differences (HANSSEN, 2001). The *Minimum Spanning Tree* ensures that the resultant set of interferograms connects all the available images with the low cost path, similar to the travelling salesman problem. The different subsets have to be treated as independent problems with this approach. In order to ensure a proper $\{B_n, T, \Delta f_{dc}\}$ distribution, we can increase the initial MST set, choosing those best interferograms which maximize the uniformity's distribution.

Different test have been performed with real data, confirming the validity of the proposed method. The resultant *MST* of the 3D image triangulation (Fig. 1(a)) of the Gardanne test site is depicted in Fig. 1(b). Here, 88 interferograms out of the initial 268 have been selected. Deformation results from this scenario will be presented in section 6.

3. Pixel Selection Criteria

The different sources of decorrelation presented have a non-uniform impact in the interferogram. Depending on the terrain characteristics, presence of edifications, changes due to human activity or natural disasters, ... some zones better preserve the coherence than others. Or in plain words, some pixels of the interferogram will present better phase quality than the others. Consequently, the differential algorithm cannot be applied to all pixels within the area under study and a selection of the reliable ones in the interferometric set has to be performed. Two different approaches are mainly employed to make an estimation of the phase quality of the pixels and base a selection criterion: the *coherence stability* and the *amplitude dispersion*.

3.1. Coherence Stability Criterion

The spatial coherence, γ, is employed under the assumption of ergodicity (HANSSEN, 2001) to obtain the maximum likelihood estimator of the coherence magnitude over an estimation window. It provides an estimation of the accuracy of the pixel's phase for each interferogram not dependent on the number of images available. The coherence estimator is defined as,

$$\gamma = \frac{\sum_{n=1}^{ML} y_1^{(n)} y_2^{*(n)}}{\sum_{n=1}^{ML} |y_1^{(n)}|^2 \sum_{n=1}^{ML} |y_2^{(n)}|^2}, \tag{2}$$

where y_1 and y_2 are the master and slave complex images, respectively. The phase of this complex coherence is in fact the multi-looked (or averaged, *ML*) interferometric

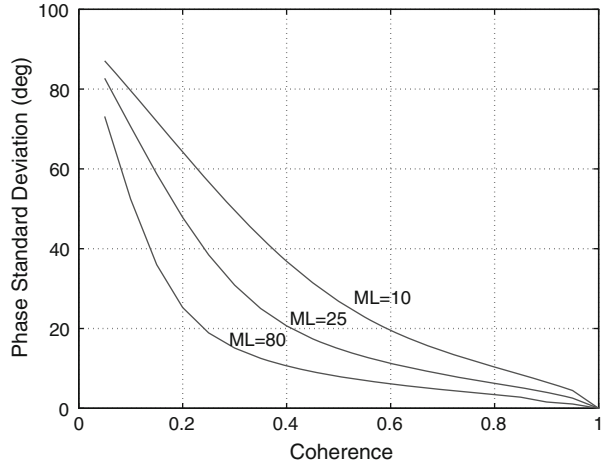


Figure 2
Phase standard deviation vs. coherence for a multi-look factor of 80, 25 and 10.

phase while its magnitude is the phase quality estimator. The size of the required coherence estimation window worsens the spatial resolution. Typical values are 3×15 pixels or 4×20 pixels (range and azimuth, respectively) which lead to a 60×60 meters and 80×80 meters ground spatial resolution. Additionally, this approach restricts the maximum perpendicular baselines up to the typical figures for classical interferometry (despite good results have been obtained with baselines up to 300 m it is recommended to keep the values as low as possible to maximize the number of pixels detected). The standard deviation of the interferometric phase, σ_ϕ, can be mathematically calculated from the estimated coherence, γ, and the multi-look (HANSSEN, 2001) as shown in Figure 2. Depending on the maximum standard deviation allowed a coherence threshold, γ_{th}, is fixed. A pixel will be selected if it presents a coherence higher than the threshold in a certain percentage of interferograms, for instance 50%.

3.2. Amplitude Criterion

This approach estimates the phase standard deviation of each pixel from its temporal amplitude stability (FERRETTI *et al*., 2001, 2002). The objective of this selection method is to find quality point-like targets, most commonly known as *Persistent* or *Permanent Scatters* (*PS*), instead of finding stable distributed targets as is done with the coherence stability method. A *PS* is a structure with an ideal isotropic response, so it is not affected by geometrical, neither Doppler centroid nor temporal decorrelation and, at least theoretically, there are no restrictions regarding the maximum spatial and temporal baselines nor Doppler centroid differences allowed. It can be demonstrated that for a high

signal-to-noise ratio the pixel's phase stability can be associated with its amplitude dispersion D_A as follows:

$$\sigma_\phi \approx \frac{\sigma_A}{m_A} \equiv D_A, \tag{3}$$

being σ_ϕ the phase dispersion, m_A the mean and σ_A the standard deviation of the amplitude values of the pixel along the image data set. The dispersion index is a good approximation of the phase stability for high SNR values. The larger the number of images the more reliable the statistical analysis is, being the desirable minimum number of scenes around 30. Typically, only those targets exhibiting a $D_A < 0.25$ (implying a mean σ_ϕ of 14°) are considered. Before the statistical analysis it is compulsory to perform the radiometric correction of the images.

As we have just commented, the usage of one selection criterion or another will determine the nature of the targets to work with. While the amplitude dispersion selects ideal point-like targets with functionable allowance at maximum spatial resolution (that provided by the SAR image), the coherence stability implies an averaging of a set of pixels, leading to a lower spatial resolution product. This averaging may also cause the loss of isolated scatters which could be detected with the amplitude dispersion criterion. Depending on the scenario, it would be more interesting to increase the number of selected points employing a coherence approach, rather than having maximum spatial resolution information. An example of this is the volcanic scenarios, where the lava characteristics do not follow a PS behavior, although attending to its spatial (large extension) and temporal (stable along time) they fit perfectly for coherence based processing. In opposition to this, we may prefer to apply amplitude processing in urban scenarios where man-made targets are more likely to be found.

Another important issue in deciding which one to employ is the number of images available from the study area. It is true that the historical archive is becoming increasing, larger, but nevertheless, we may just get a few images solely because there was a poor image disposability in the temporal period in which we are interested. If having a low number of them, the D_A estimator is not reliable. On the other hand, the coherence estimator is more robust when dealing with a low number of interferograms (CPT has obtained good results employing seven images under this criterion). Consequently, in this situation we may be more restrictive on the D_A estimator (lower threshold values) however this may lead to a lower density of selected pixels, which is not at all desirable, as the larger the spatial sampling among the set of interferograms, the better the parameters estimation. Therefore, a coherence based selection will be preferred here. For both criterions there is always a compromise between the number of pixels selected and its reliability.

In order to perform multi-layer processing (which is explained in the following section), selected pixels can be divided in different layers according to their quality.

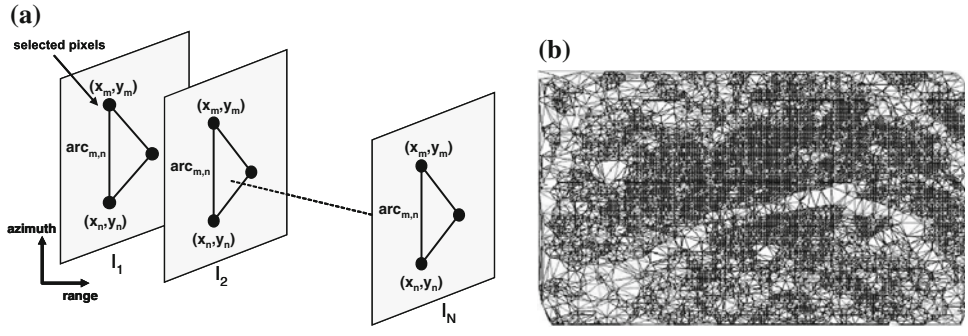


Figure 3
(a) Example of Delaunay triangulation: Two nodes define an arc (b) triangulation of the selected pixels in the city of Paris where darker colour means higher arc quality.

4. CPT Linear Block

The phase of individual pixels is difficult to use due to the presence of different phase offsets among the set of differential interferograms. Those offsets can be calculated over high coherence stable areas not affected by deformation and atmospheric artefacts, but in this case additional input information would be required. In CPT this problem has been overcome by relating the neighboring selected pixels by means of a Delaunay triangulation and cancelling then the offset effect. This kind of triangulation relates all the neighboring pixels of irregularly gridded data generating non-overlapped triangles, as shown in Figure 3. The selected pixels are the *nodes* of the grid and each pair defines an *arc* of the triangulation.

In order to reduce the effects of the atmospheric artefacts on the estimation of the linear component of deformation, a maximum arc length is set when performing the Delaunay triangulation. This maximum distance between connected nodes depends on the assumed value for the correlation distance of the atmosphere, usually 1 km. Nevertheless, sometimes this restriction can be relaxed to avoid the existence of disconnected areas.

After the triangulation, the phase increment between two neighboring pixels can be expressed as

$$\begin{aligned}\Delta\phi^k(T_i, B_n, \Delta f_{dc}, x_m, y_m, x_n, y_n) &= \Delta\phi^k(T_i, B_n, \Delta f_{dc}, arc_{m,n}) \\ &= \phi^k(x_m, y_m) - \phi^k(x_m, y_n),\end{aligned} \tag{4}$$

where k stands for the sensor (ERS or ENVISAT), (x_m, y_m) and (x_n, y_n) are the coordinates of the nodes forming the $arc_{m,n}$. If the maximum connecting distance between pixels has been properly fixed, then the atmospheric term can be eliminated from equation (4). Since the linear velocity term and DEM error are constants for each node (and arc) in the whole set of differential interferograms, it is possible to retrieve a good estimation of them adjusting the phase model, $\Delta\phi^k_{model}$, to the available data, $\Delta\phi^k$.

$$\Delta\phi^{k}_{model}(T, B_n, \Delta f_{dc}, arc_{m,n}) = \frac{4\pi}{\lambda_k} \cdot T \cdot \Delta v(arc_{m,n}) + \frac{4\pi}{\lambda_k} \frac{B_n(T)}{r(T)\sin\theta} \cdot \Delta\varepsilon(arc_{m,n}) - \frac{2\pi}{v_s} \Delta f_{dc} \cdot \Delta az(arc_{m,n}), \tag{5}$$

The last term of equation (5) is only applied when working with *Permanent Scatters*. In order to find the values Δv, $\Delta\varepsilon$ and Δaz for each arc, a test function is minimized. This function is called the *model adjustment function* $\Gamma(arc_{m,n})$ and is defined as,

$$\Gamma(arc_{m,n}) = \sum_{i=1}^{N_{ERS}} \left| \exp[-j(\Delta\phi_i^{ERS}(T, B_n, \Delta f_{dc}, arc_{m,n}))] - \exp[-j(\Delta\phi_{i,model}^{ERS}(T, arc_{m,n}))] \right|^2 + \sum_{i=N_{ERS}+1}^{N_{ERS}+N_{ENV}+1} \left| \exp[-j(\Delta\phi_i^{ENV}(T, B_n, \Delta f_{dc}, arc_{m,n}))] - \exp[-j(\Delta\phi_{i,model}^{ENV}(T, arc_{m,n}))] \right|^2, \tag{6}$$

where N_{ERS} and N_{ENV} are the number of ERS and ENVISAT interferograms, respectively, so their addition conforms to the total number of employed interferograms N. The minimization of Γ is equivalent to find the bidimensional (tridimensional if accounting for the azimuth target's *subpixel* position) frequency of the complex sinusoid derived from the phase term in equation (5). The Γ minimization is done applying the Conjugate Gradient Method (CGM) (Press *et al*. 2002). As the minimization is done in the complex plane, there is no need to perform any kind of phase unwrapping on the interferograms. In fact, with a good distribution of spatial and temporal baselines it is possible to derive the linear terms of the model for an arc even when its interferometric phases were wrapped in all interferograms. The larger the number of interferograms, the better will be the estimation as the impact of atmospheric artefacts and decorrelation noise is reduced. There is no clear minimum of images as results depend on the particularities of each case, but CPT obtained successful results even with only seven images (Mora *et al*., 2003). The noise sensitivity of the retrieved parameters depends on the range and distribution of baselines available. The larger the spatial baselines, the more precise will be the estimation of the DEM error. Similarly, uniform distributions of baselines provide better results. The minimization of Γ has to be done carefully as it presents several local minima.

Once the minimization process has been accomplished for each arc, the result is the following set of velocity, DEM error and azimuth position increments.

$$\begin{aligned} \Delta\tilde{v}(x_m, y_m, x_n, y_n) &= \tilde{v}(x_m, y_m) - \tilde{v}(x_n, y_n) \\ \Delta\tilde{\varepsilon}(x_m, y_m, x_n, y_n) &= \tilde{\varepsilon}(x_m, y_m) - \tilde{\varepsilon}(x_n, y_n) \\ \Delta\tilde{az}(x_m, y_m, x_n, y_n) &= \tilde{az}(x_m, y_m) - \tilde{az}(x_n, y_n). \end{aligned} \tag{7}$$

The minimization of Γ always provides a pair of velocity and DEM error values, even if the phases were random values. Consequently it is necessary to evaluate the quality of

the solutions to discard those which are incorrect. The *model quality function*, γ_{model} is used as a trustworthiness test to discard those arcs with non-reliable values. The γ_{model} is defined for each arc as follows,

$$\gamma_{model}(arc_{m,n}) = \frac{1}{N}\left|\sum_{i=1}^{N_{ERS}} \exp[-j(\Delta\phi_i^{ERS}(arc_{m,n}) - \Delta\phi_{i,model}^{ERS}(arc_{m,n}))] + \sum_{i=N_{ERS}+1}^{N_{ERS}+N_{ENV}+1} \exp[-j(\Delta\phi_i^{ENV}(arc_{m,n}) - \Delta\phi_{i,model}^{ENV}(arc_{m,n}))]\right|, \quad (8)$$

which is equal to one if the model perfectly fits the data, and tends to zero with total decorrelation. Those arcs presenting a quality function below a threshold will be rejected. Here it is necessary to consider the possibility of having a nonlinear movement term in the phase increment, as it is a deterministic term and we risk considering it as noise. For this reason, the fixed threshold $\gamma_{model\ th}$ is set to a lower value than that corresponding to a linear deformation pattern.

Finally, an integration process is necessary to obtain the absolute values for each pixel. As the solution is obtained from the relations among the nodes, it is necessary to include at least a control point of known linear velocity and height. In practice, many stable points not affected by deformation are used, depending on the extent of the area and the in-field information available. A good distribution of control points helps to reduce the offsets that could appear among zones badly connected. CPT performs the integration using the *Conjugate Gradient Method* (*CGM*), an iterative and efficient method used for solving large systems of linear equations (SARKAR *et al*., 1984).

Depending on the chosen coherence or amplitude threshold, selected pixels may vary in a wide quality range, and consequently so it will the quality of the estimations on the increments of the linear parameters. Furthermore, wrong arc estimations will lead to wrong absolute values after the integration process. In order to maximize pixel density but preserving result quality *multi-layer processing* has been implemented (BLANCO *et al*., 2006). Selected pixels are divided in to different layers according to their quality (expressed by their coherence or amplitude dispersion value). After that, beginning with the top layer, the linear block (depicted in Fig. 4(a)) is iteratively executed by adding successive layers, so the obtained absolute values of each layer act as the seed values to the following integration process. By doing this, the results obtained with the high quality layers are preserved and the estimation of the low quality layers improves. Consequently, multi-layer processing improves linear results and rises pixel density while providing a *quality label* for each one. Figure 5 illustrates this process for a simple 3 quality layers case. As it is seen, the linear block is iteratively applied adding quality layer pixels (from best quality black to lowest quality blue pixels) at each iteration. At the end of each iteration the absolute velocity and DEM error values of the corresponding layer pixels are fixed, serving as seed values in the following iteration integration process.

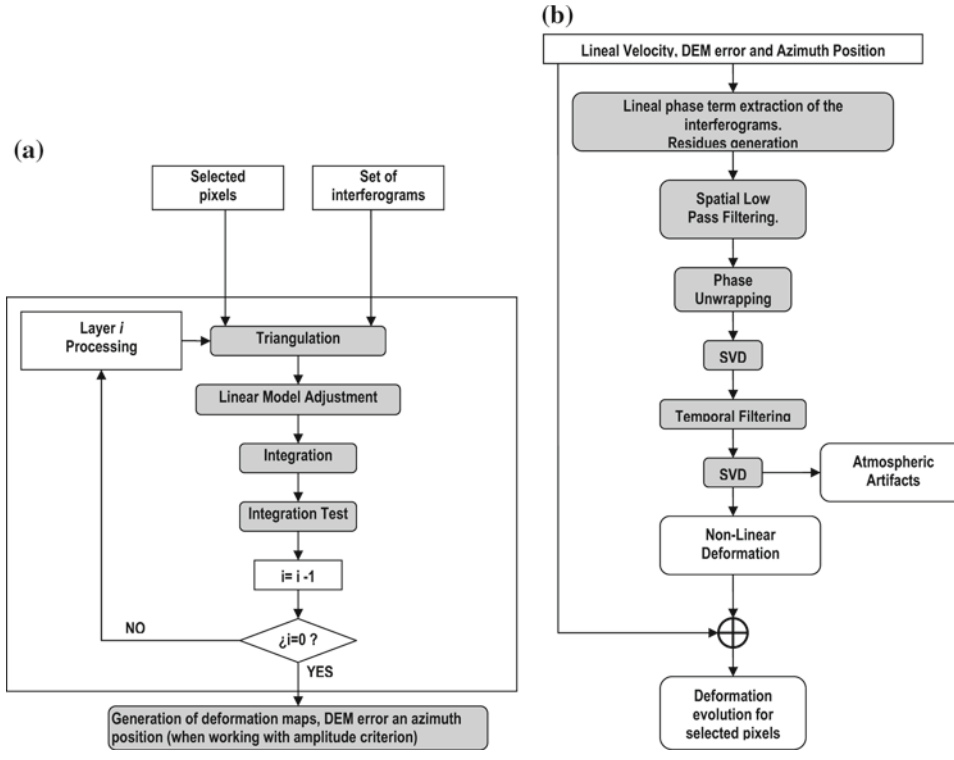


Figure 4
(a) CPT Linear block and (b) Nonlinear block layout.

5. CPT Nonlinear Block

After calculating the linear deformation map, it is possible to obtain the nonlinear component to complete the study of displacement. Adding this nonlinear deformation component to the linear term, a detailed plot of the evolution of deformation is obtained. The first step of the nonlinear block of CPT, whose layout is shown in Figure 4(b), is the calculation of the phase residues $\phi^k_{residue}$, obtained by subtracting the absolute linear phase model $\phi^k_{model}(\tilde{v}, \tilde{\varepsilon}, \tilde{az})$, from the original interferometric phases ϕ^k for all selected pixels,

$$\phi^k_{residue}(T, x, y) = \phi^k(T, x, y) - \phi^k_{model}(T, x, y), \tag{9}$$

where (x, y) stands for the selected pixel coordinates. The linear phase model is obtained from the estimated linear terms with,

$$\phi^k_{model}(T, x, y) = \frac{4\pi}{\lambda_k} \cdot T \cdot \tilde{v}(x, y) + \frac{4\pi}{\lambda_k} \frac{B_n}{r \sin\theta} \cdot \tilde{\varepsilon}(x, y) - \frac{2\pi}{\nu} \cdot \tilde{az}(x, y). \tag{10}$$

The residual phases are calculated only on the reliable pixels that have survived the different quality tests. After this step, the phase residues can be expressed as

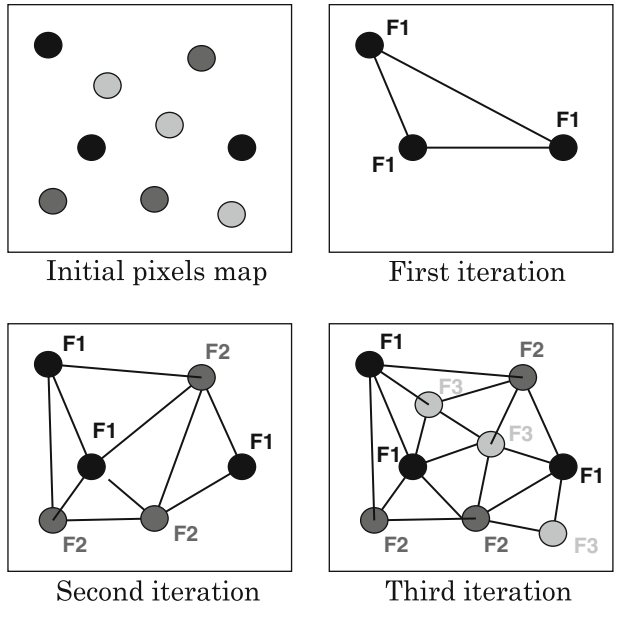


Figure 5
Selected pixels map, best quality pixels in black, middle quality pixels in red and low quality pixels in blue. First iteration where the linear block is applied to black pixels, second iteration where the linear block is applied to black+red pixels and third block where the linear block is applied to black+red+blue pixels.

$$\phi^k_{residue}(T,x,y) = \phi^k_{atmos}(T,x,y) + \phi^k_{nonlinear}(T,x,y) + \phi^k_{noise}(T,x,y), \qquad (11)$$

where two essential phase terms have to be considered: the atmospheric artefacts, ϕ^k_{atmos}, and the nonlinear displacement, $\phi^k_{nonlinear}$. Both can be separated, taking advantage of their different frequency characteristics in space and time:

- **Atmospheric perturbations**. This term can be considered as a low spatial frequency signal in each image due to its approximately 1 km correlation distance. However for each acquisition date atmospheric conditions can be considered random and, consequently, the atmospheric contribution can be modelled as a white process in time.
- **NonLinear deformation**. On the other hand, this term can be assumed to present a narrower correlation window in space (or at least much narrower than atmospheric artefacts) and a low-pass behavior in time.

Taking into account all these considerations, the separation of the atmospheric artefacts from the nonlinear deformation can be implemented with a filtering process in both spatial and time domains.

Firstly, a spatial low-pass filtering using a two-dimensional moving averaging window of 1 km $\times$ 1 km (typical correlation distance of atmosphere (HANSSEN, 2001) is applied to the residual phase, $\phi^k_{residue}$. After this step two components should remain:

$$\phi^k_{resSLR}(T,x,y) = \phi^k_{atmos}(T,x,y) + \phi^k_{nonlinearSLR}(T,x,y), \tag{12}$$

where $\phi^k_{res\ \ SLR}$ is the nonlinear component of the displacement at *Spatial Low Resolution* (*SLR*) and ϕ^k_{atmos} the atmospheric artefacts, which can be assumed not to be affected by the spatial filter. The *SLR* residue is composed of the atmospheric perturbations, which are signals with a low variation in space, and the spatial low resolution version of the nonlinear displacement, which behaves in space similarly to the atmospheric component. Some authors (BERARDINO *et al.*, 2002) adopt another spatial filtering strategy which consists in doing it on the phase image contribution (φ in equation 13) instead of doing it on the interferograms as CPT does. As it is explained in the following, the image phase contribution (φ eq.13) is calculated employing the Singular Value Decomposition (SVD), which needs the interferometric phases to be unwrapped. As these techniques have already unwrapped the phase up to this point, the SVD can be applied on the interferometric phases, performing afterward the spatial filtering on the image phase contribution, so the number of filtering operations is reduced because the atmospherical interferometric phase component is a lineal combination of the atmospherical image phase contribution. CPT unwraps the phases at the present processing step (i.e., previously to image phase contribution), so the heavily filtered interferograms resultant of this spatial filtering $\phi^k_{res\ \ SLR}$ are especially easy to unwrap as they are usually very smooth and contain almost no fringes. At this step the Conjugate Gradient Method is employed to unwrap the residual phases, similar to the integration step.

The interferograms have been formed from the phases of two SAR images separated in time and they do not follow the temporal order required by the temporal filter that has to be applied. The interferometric phase can be expressed in the function of their forming images,

$$\phi^k_{resSLR}(T,x,y) = \varphi^k_{resSLR}(TM,x,y) - \varphi^k_{resSLR}(TS,x,y), \tag{13}$$

where *TM* and *TS* are respectively the acquisition time of *master* and *slave* images. The *Singular Value Decomposition* (*SVD*) is used to obtain the phase of each image from the stack of interferograms. Once the absolute temporal information (with respect to the first image, strictly speaking) has been obtained, a high-pass temporal filter can be applied in order to estimate the atmospheric contribution and isolate the nonlinear deformation at low spatial resolution. In this step a triangular filter is employed with a fixed temporal span. Talking in frequency terms, the main difficulty resides in the selection of the cutoff frequency to discriminate the atmospheric component from the nonlinear displacement, as the atmospheric effects appear in all frequencies while the nonlinear displacement is expected to be a low-pass signal. Therefore, the cut frequency should be placed on the considered highest frequency for the nonlinear displacement. After the temporal filtering, the phase residue should contain only the *SLR* nonlinear deformation, $\varphi_{non-linearSLR}$.

The obtained deformation is not complete, because it has been calculated using spatial low-pass filtered phases. The procedure to obtain the remaining *Spatial High Resolution* (*SHR*) component is similar to the one described but starting from a better phase model using all the available information, including the estimated linear and *SLR* nonlinear deformation, the DEM error and the estimated atmospheric artefacts. Consequently, a new *SHR* residue can be obtained by subtracting the improved model ϕ'^{k}_{model} from the original interferometric phases. Then, this new residue is basically composed of two terms:

$$\phi^{k}_{resSHR} = \phi^{k} - \phi'^{k}_{model} = \phi^{k}_{nonlinearSHR} + \phi^{k}_{noise}, \tag{14}$$

and from here forward the same *SLR* estimation schema (unwrapping+SVD+temporal filtering) is followed to obtain the *SHR* nonlinear movement term. The temporal deformation evolution is obtained by adding their three estimated components, i.e., the linear deformation term and the low and high spatial resolution nonlinear deformation terms,

$$\rho(t_i) = \tilde{v}_{est} t_i + \frac{\lambda_k}{4\pi}\left[\varphi^{k}_{non-linearSLR}(t_i) + \varphi^{k}_{non-linearSHR}(t_i)\right]. \tag{15}$$

This deformation has been calculated on the so-called *line-of-sight* or *slant-range* direction. The vertical deformation can be obtained using the local incidence angle θ,

$$\rho_{vert}(t_i) = \frac{\rho(t_i)}{\cos\theta}, \tag{16}$$

6. Results

6.1. DEM Extraction

In order to test the extraction of the topography with CPT, we have processed the Paris data set without cancelling its topography, consequently the calculated DEM error should be the true topography of the scenario. The spatial and temporal baseline distribution of the Paris set of interferograms is depicted in Figure 6. In it, each circle represents the employed images while the arcs connecting them represent the employed interferograms. The resulting set calculated out of 62 images (48 ERS and 14 ENVISAT) has been achieved by the explained MST selection and is composed of 78 interferograms with a maximum spatial baseline of 300 m, a maximum temporal baseline of 1000 days and a maximum Doppler centroid difference of 500 Hz.

The results, obtained with different maximum spatial baselines (50 and 150 meters) are compared to a 10 meter resolution DEM. Mean and standard deviation values of the error are shown in Table 1. As expected, the higher the baseline the better the results, but in any case, with a maximum standard deviation value of 4 meters, topography is well extracted, as we can see in Figure 7.

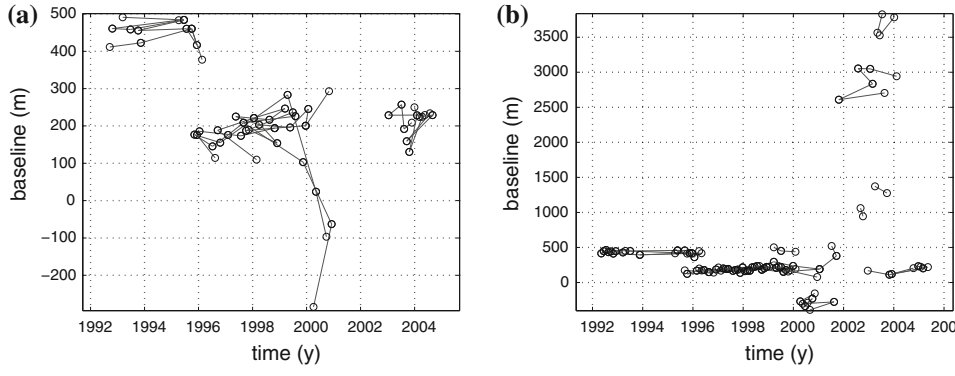


Figure 6
Spatial and temporal baseline distribution of (a) Paris and (b) Gardanne.

Table 1

Mean and Stdv of DEM real - DEM calculated

Maximum Baseline	Mean (DEM-cal.DEM)	Stdv (DEM-cal.DEM)
50 m	7.59 m	4.15 m
100 m	4.41 m	3.37 m
150 m	4.34 m	3.30 m

Mean and Standard deviation values of the error of the calculated DEM error when employing different data sets with different maximum spatial baselines

6.2. Linear Deformation

In order to show the performances of CPT when dealing with linear type deformation movements, three different scenarios have been analyzed. The first one is the city of Murcia (SE Spain). Here, long drought periods and the over-exploitation of their water infiltration aquifer reserves have provoked a deformation episode. Figure 8 shows the linear velocity map of the area obtained by CPT where four points have been selected to display its temporal deformation. These results were obtained by the University of Alicante (UA) in collaboration with the Instituto Geologico Minero de España (IGME) and UPC (Tomas *et al.*, 2005). The employed data set constitutes 47 images (36 ERS and 11 ENVISAT) and 83 interferograms with a maximum baseline of 100 m, a maximum temporal baseline of 1000 days and a maximum Doppler centroid difference of 300 Hz. As we can see, the deformation pattern of the area is highly linear.

In Figure 9 an estimation of the land settlements of the same area of Murcia is displayed, where the four selected points are also located. This estimation has been carried out through theoretical-mathematical modelization considering land characteristics and the drawing down of the water level (Mulas *et al.*, 2003). By comparing both DInSAR and theoretical results we want to show, rather than absolute deformation values, the high correlation between both spatial deformation gradients.

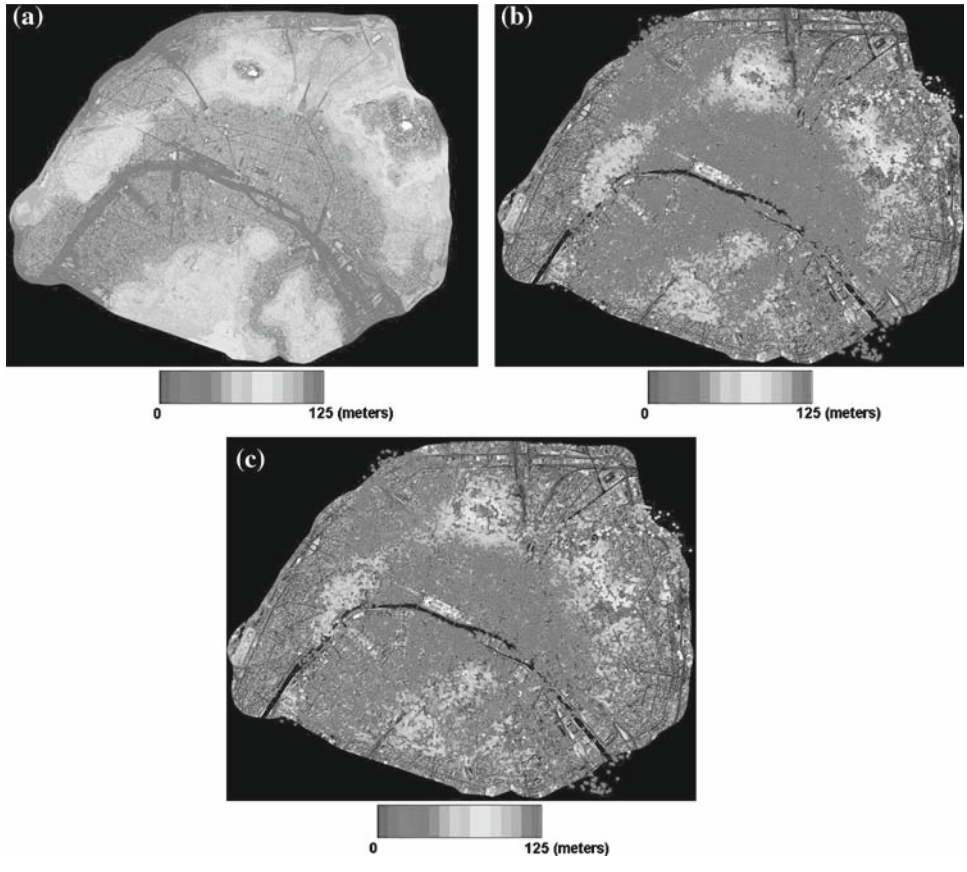


Figure 7
(a) DEM of the city of Paris, (b) calculated height map of the same area for the selected pixels employing an interferogram set with a maximum baseline of 50 meters and (c) 150 meters.

When doing so, it is possible to observe that there is a high correlation between the estimated movement and the theoretical prediction when moving from different colored areas. It is also interesting to comment that these results have been obtained with a low number of images. Evidently, a linear deformation movement requires a temporal sampling less demanding than a nonlinear pattern. In reference to the presented study, a further study of the Metropolitan area of the city of Murcia has recently been carried out. Results, concerning validation between CPT DInSAR series and ground truth data are resumed in a forthcoming publication submitted to the Journal of Photometry and Remote Sensing (HERRERA *et al*., 2008). Comparison between CPT and extensometers deformation series reveals differences with standard deviation values around 2.3 mm.

The second test site is the neighborhood of Mont-Martre in Paris. The employed data set is the one depicted in Figure 6(a). This scenario was a mining area whose gypsum

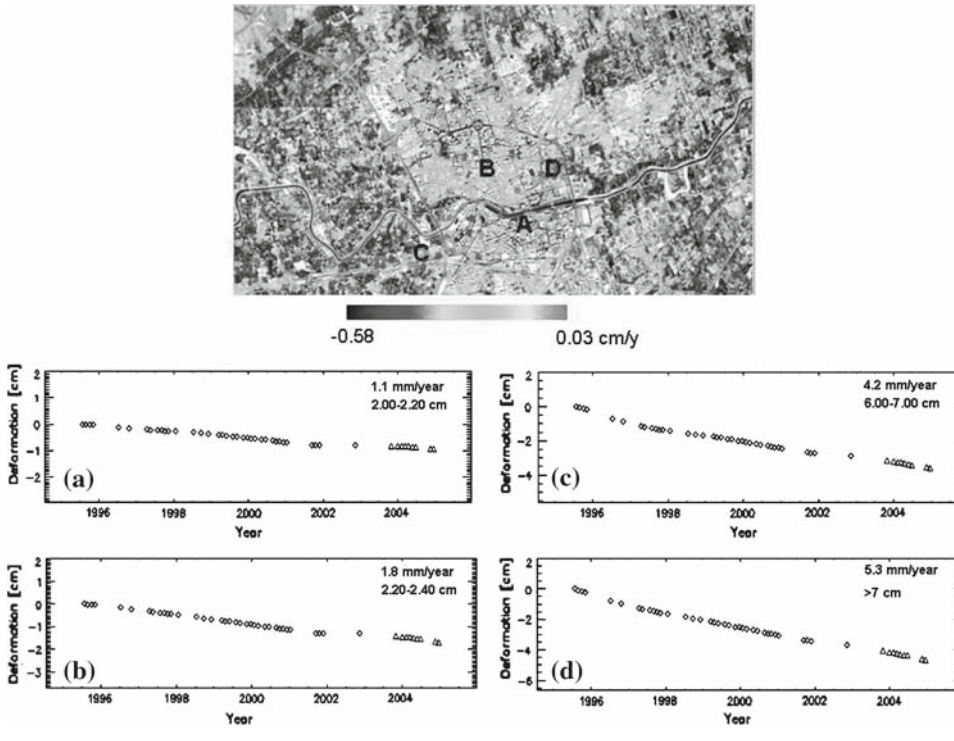


Figure 8
Linear velocity map of the area of the city of Murcia where 4 points have been selected. Temporal deformation series of the selected points with their corresponding theoretical deformation depicted in Figure 9.

mines were improperly filled once the activity finished. The gypsum dissolution and some slope movements cause deformation phenomena, particularly specially in the hill of Mont-Martre. If paying attention to the linear deformation map of this area (Fig. 10), the hill is delimited by the darker coherent points which present a subsidence velocity. Furthermore, the deformation evolution with coherence and amplitude method (left and right side, respectively) for 3 points, 2 located on the hill (A and B) and one in the stable area (C), are represented. As it is seen, the expected linear deformation behavior is correctly reproduced with good agreement between amplitude and coherence results. In order to better interpret a comparison between coherence and amplitude results, it is important to remind that both criteria select different kinds of targets so their spatial resolution is different. Nevertheless, as selection criteria in terms of phase standard deviation have been similar, it is normal to achieve similar results.

The authors did not have access to ground-truth data of this area so no direct evaluation of the CPT temporal series was performed. Nevertheless, CPT results show great agreement to those presented in diverse articles (DEFFONTAINES *et al.*, 2004; SIMONETTO *et al.*, 2005) where this area was analyzed.

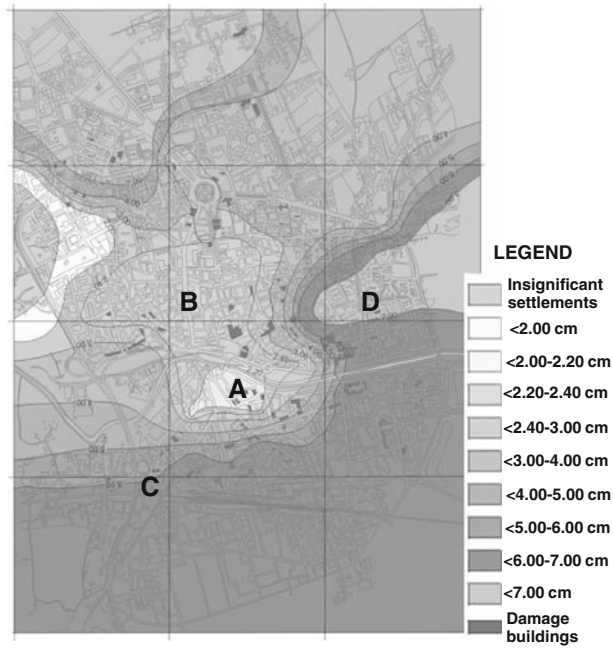


Figure 9
Deformation results obtained by MULAS *et al*. (2003) using piezometric measurements and mathematic models. The same four selected points in Figure 8 have been located.

The third test site is the city of La Union (SE, Spain), which is located within a metal mining area that has been exploited since the Roman Period, with a recent activity peak in the second half of the 19th and throughout the 20th century until the 1980s. This historic exploitation has left behind a high concentration of underground cavities due to the abandoned mine galleries. This work was also carried out by UA, in collaboration with the IGME and the UPC (HERRERA *et al*., 2007). In order to validate the CPT results in this scenario, they have been compared to topographical leveling data. The topographical leveling is comprised of 57 measurement points located towards the SW of La Union, deployed in two measuring campaigns performed in April 2003 and July 2004 (CIMA, 2005). On the other hand, 14 images (6 ERS and 8 ENVISAT), acquired from April 2003 to 2004, forming 41 interferograms with a maximum spatial baseline of 150 m, a maximum temporal baseline of 900 days and a maximum Doppler centroid difference of 300 Hz., have been processed with CPT.

Figure 11 shows the absolute deformation for the CPTs selected pixels as well as for the topographical leveling measurement points. In it, pixel *a* is the DInSAR reference control point while *P1* is the control point for the topographical leveling network. The DInSAR retrieved deformation value in the correspondent pixel of *P1* is -0.1 cm, which confirms the good correlation between both data sets seen in Figure 11. Deformation of the

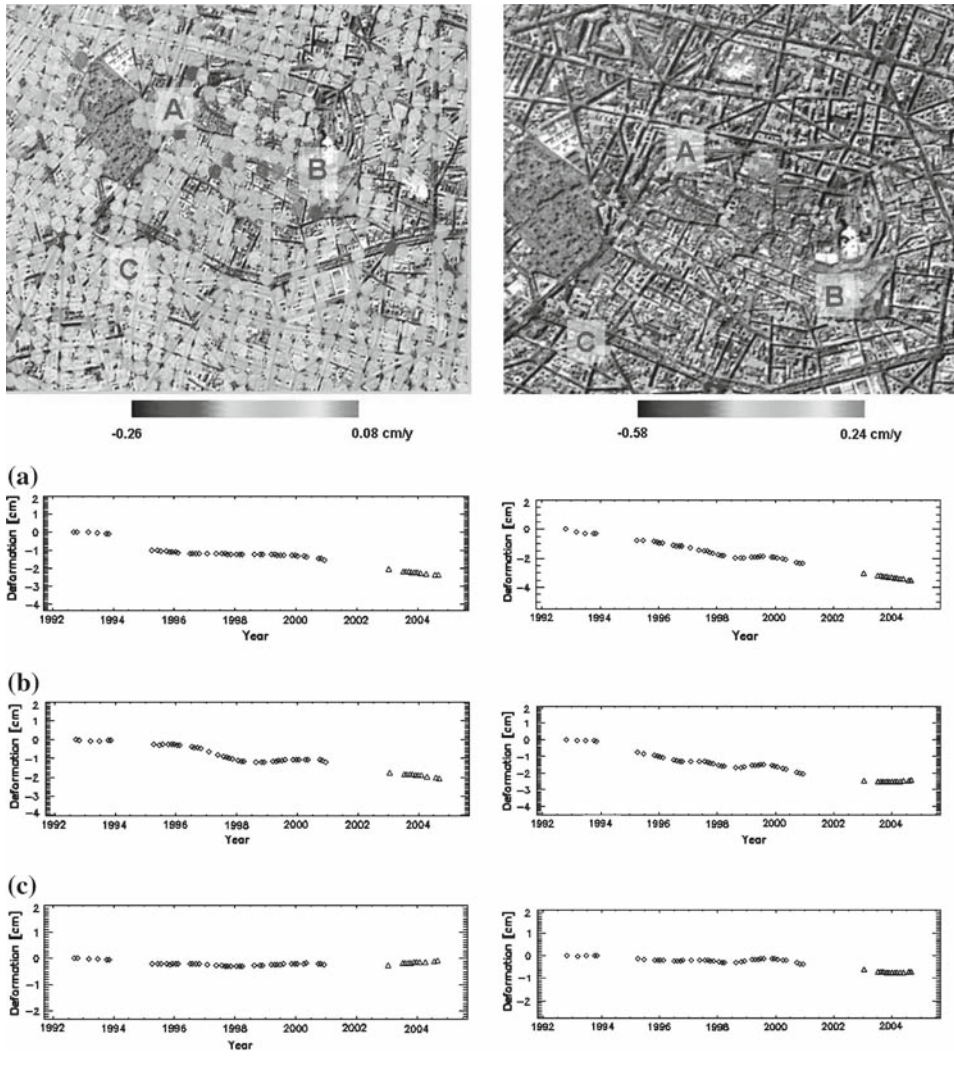


Figure 10
Linear velocity map of the Mont-Martre area (Paris) where some points have been selected and their respective deformation graphs (left) when using the coherence stability pixel selection criterion (3 × 15 ML factor) (right) when using the amplitude dispersion criterion.

coincident pixels has been deployed in Figure 12 for comparison purposes. As is seen, the absolute differences vary from 0 to 1.9 cm with an absolute mean difference of 0.7 cm and the standard deviation is 0.5 cm. In addition to this, it is important to mention that as the DInSAR pixels have a size of 80 × 80 m and some ground control points are located on the same pixel. Taking into account the most similar deformation value of the ground control points with respect to the pixel they are included in, the absolute mean difference falls to 0.5 cm while the standard deviation falls to 0.3 cm.

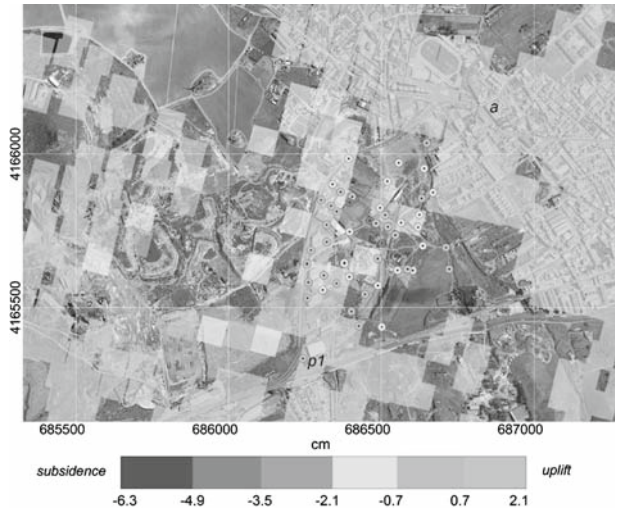


Figure 11
DInSAR and topographical leveling deformation in the SW area of La Union.

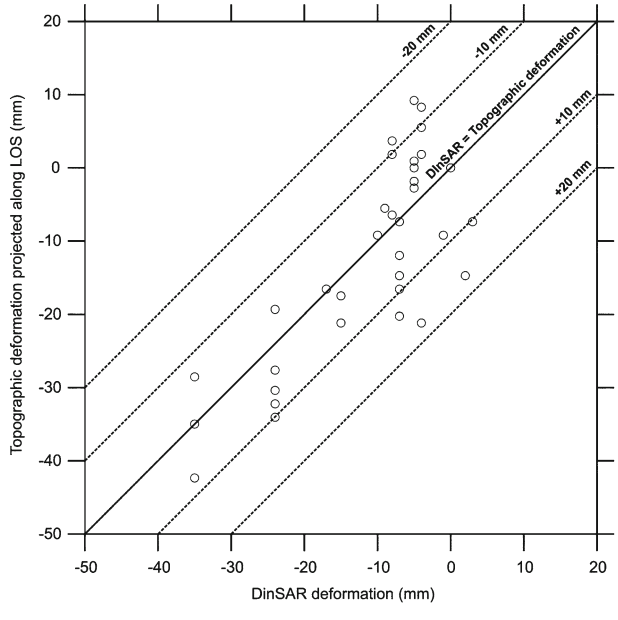


Figure 12
Comparison between CPT estimated deformations and topographical leveling measured deformations projected along the Line of Sight (LOS).

6.3. Nonlinear Deformation

We now show the results obtained over areas which present a nonlinear deformation movement pattern. Two scenarios have been selected, St. Lazare Railway Station in Paris, and Gardanne, located in the South-East of France. Figure 13 shows the results corresponding to the St. Lazare Railway station where underground works caused a deformation movement. The employed data set is the one depicted in Figure 6(a). Again, three points have been selected in the linear velocity map for showing their temporal deformation. This movement is basically composed of a subsidence movement caused when water was extracted in order to keep the works dry, a stability period and an uplift period when the phreatic level was restored. On the left side the temporal deformation obtained with the coherence method is depicted, while those with the amplitude criterion are on the right side. For comparison purposes, both results have been obtained with the interferometric data set in Figure 6(a). Nevertheless, more interferograms could have been used in the amplitude case, as the baseline restrictions can be relaxed. As we can see, the deformation plots are very similar in both methods and the expected deformation pattern is well reproduced in form and in magnitude, yielding a maximum subsidence value around -2 cm.

Even though the authors did not have direct access to ground-truth data of this area, CPT results show high matching to those present in diverse articles where this area was analyzed (FRUNEAU *et al*., 2003 and 2005), carrying ground-truth data comparison.

The last results to presented correspond to the area of Gardanne (France). The spatial and temporal baseline distribution of the Gardanne set of interferograms is depicted in Figure 6(b). The resulting set calculated out of 97 images (89 ERS and 8 ENVISAT) has been achieved by the explained MST selection and is composed of 91 interferograms with a maximum spatial baseline of 300 m, a maximum temporal baseline of 1000 days and a maximum Doppler centroid difference of 500 Hz. Mining activity in this area has provoked a nonlinear subsidence movement whose changing rate is related to the exploitation periods of the different mines. In Figure 14, the linear velocity map of the total studied area is depicted, showing a very localized area where a subsidence movement is occuring. Another area which presents stability has also been selected for detailed study (both marked in yellow). Their corresponding velocity maps are displayed in Figure 15. Different points have been selected on both maps in order to represent their temporal deformation series. In Figure 15, 6 points are marked A-F, 5 points located on the deformation area and 1 on the stable area. A general nonlinear pattern can be distinguished in these plots. It consists of an initial slow subsidence movement (even initial stability has been detected) followed by a rapid falling in a short period of time (approximately a period of two years) to finally re-establish stability. The collapse instant varies depending on the area where the pixel is placed as does the beginning of the final stability. In the stable area one point has been selected.

This area has served as the test site of an intercomparison project promoted by ESA called PSIC-4 where several DInSAR algorithms (were the CPT was also included)

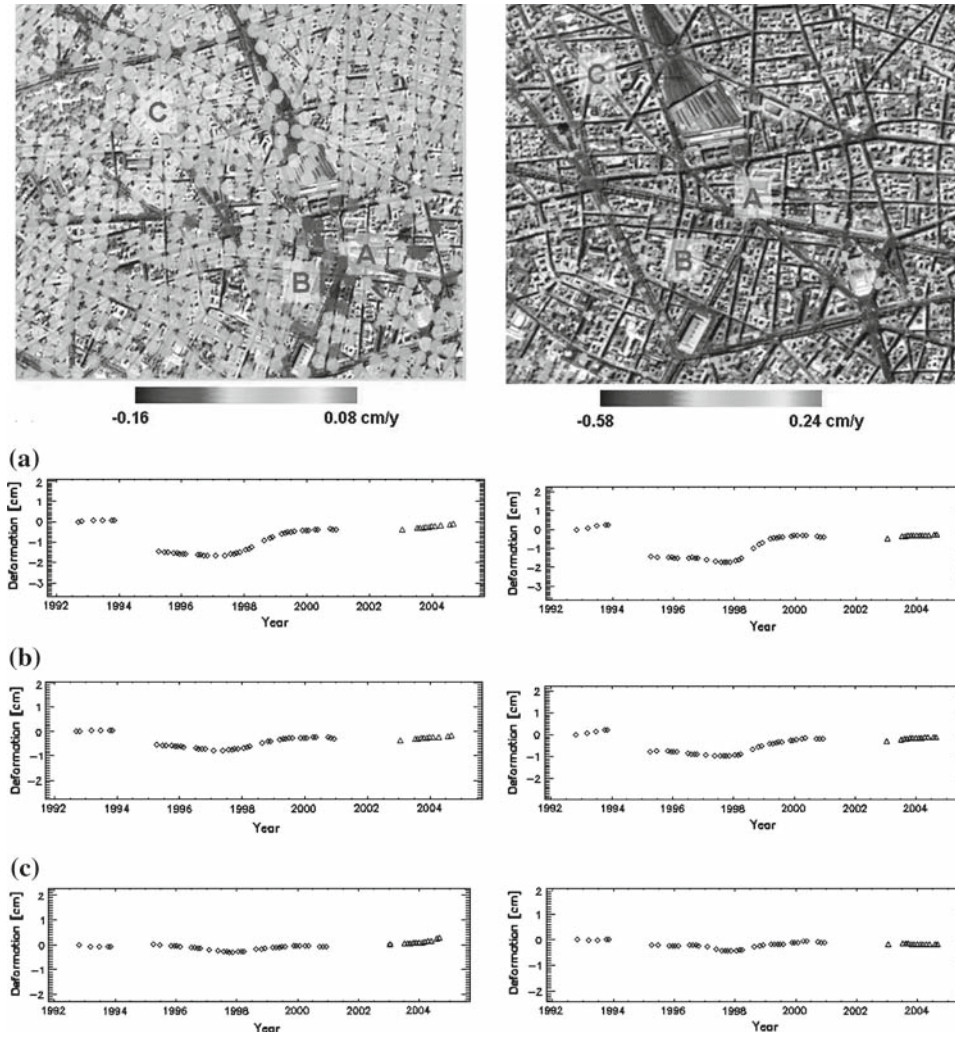


Figure 13
Linear velocity map of the St. Lazare Railway Station (Paris) where some points have been selected and their respective deformation graphs (left) when using the coherence stability pixel selection criterion (3 × 15 ML factor) (right) when using the amplitude dispersion criterion.

performances were compared to ground-truth data of the area. Documentation related to this project containing ground-truth data can be found in ESA (2007).

7. Conclusion

In this paper we have presented the potentials of DInSAR interferometry as a valuable tool for detecting and monitoring deformation movements as well as for furthering

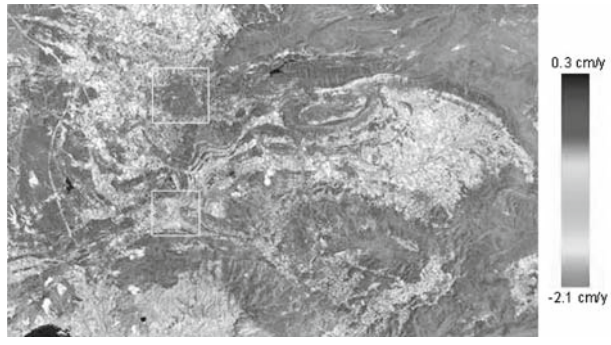


Figure 14
Gardanne linear velocity map. Highlighted white box areas correspond to a stable are (upper) and a subsidence (lower) area which are selected for a detailed study in Figure 15.

understanding of the geological mechanisms behind them. In particular, the CPT DInSAR algorithm has been described pointing out their main characteristics as well as presenting results over different areas. The presented results demonstrate the ability of CPT (which is generally extendable to multiple images DInSAR techniques) to work and retrieve deformation under a wide range of conditions (multi-looked or single-look data, linear and nonlinear deformation patterns, urban and non-urban scenarios...).

Generally speaking of DInSAR techniques, an interesting DInSAR accuracy assessment can be found in LANARI *et al*. (2007a and 2007b), FERRETTI *et al*., (2007) revealing the power of these techniques to offer millimetric accuracies. CPT accuracies have been mentioned whenever ground-truth data were available for comparison purposes. Typical figures reveal millimetric precisions as described for the Vega Media and La Union. Nevertheless, further efforts must be undertaken between the DInSAR developers and the final users of these techniques in order to set ground-truth validation campaigns so more conclusive CPT accuracy numbers will be given. A qualitative comparison in which no ground-truth data were available has been provided, confirming the accuracy revealed by *in situ* measurements as being comparable to other DInSAR techniques.

Actual efforts are brought about to improve the existent algorithms by introducing more sophisticated models in the differential processing with the aim of achieving better accuracies. Interaction with the geophysical community also will be extremely important to perform a critical evaluation and a notable development of these techniques. The upcoming new SAR sensors (TerraSAR-X, ALOS-PALSAR, RADARSAT-2) will help to increase temporal monitoring as well as spatial resolution. The exploitation of different microwave bands with these new sensors may also register an increase of the DInSAR capabilities in other environments as land-scape scenarios. Furthermore, the new future acquisition configurations as the ground-based SAR (PIPIA *et al*., 2007) and the bistatic SAR (SANZ-MARCOS *et al*., 2007) will aid study of deformation phenomena, avoiding the

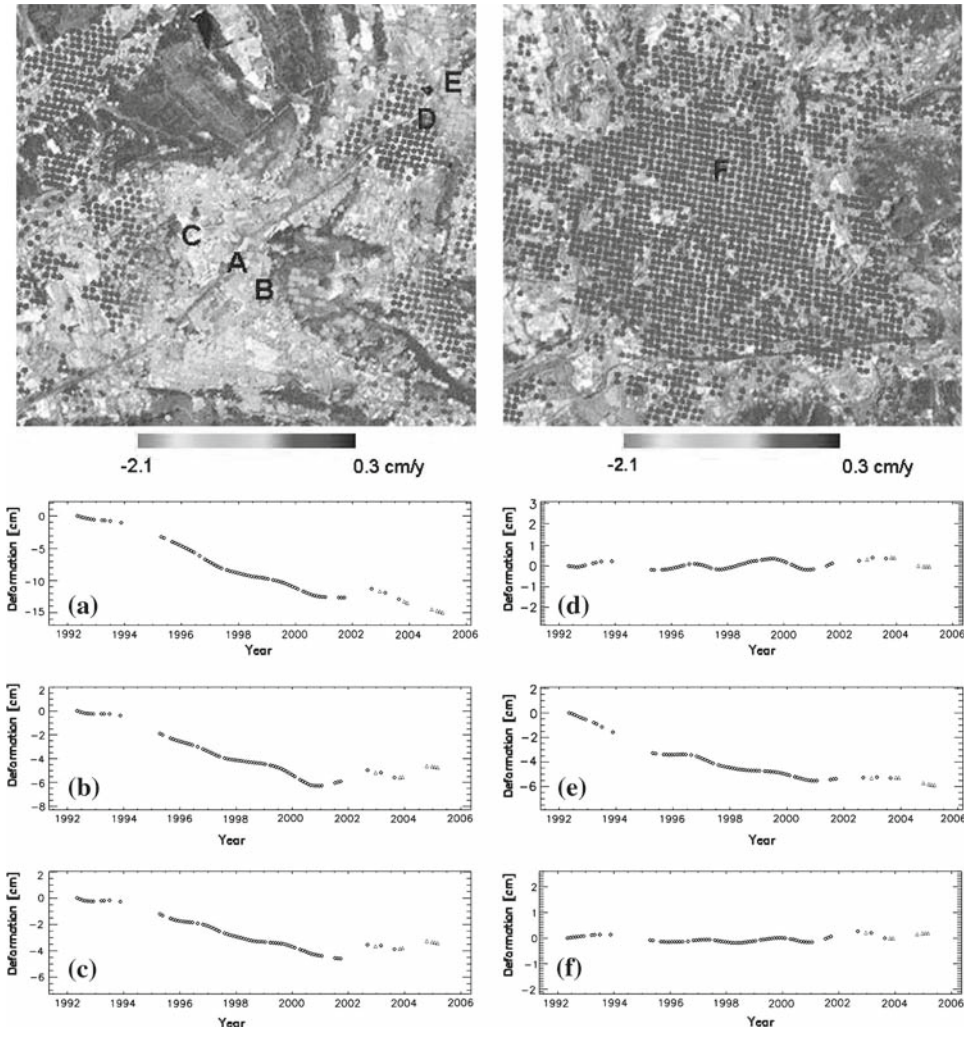


Figure 15
Gardanne's subsidence (left) and stable (right) areas linear velocity maps (highlighted in Fig. 14) and the corresponding deformation series of the selected points.

inherent space-borne geometrical limitations, adjusting their configuration according to the particular scenario to be studied.

Acknowledgements

This work has been sponsored by ESA under projects "Development of Algorithms for the Exploitation of ERS-ENVISAT Using the Coherent Pixel Technique", "PSIC4

(Persistent Scatters Interferometry): Processing Over a Validation Test Site" (Contract 18708/04/I-LG), CAT1-3148 and CAT1-1195. This work has also been financed by the Spanish MCYT under the project TEC2005-06863-C02-01. The NASA/JPL/DLR SRTM mission has provided DEMS. The authors would like to thank R. Tomás and J. M. Sánchez-López from the UA and G. Herrera and J. Mulas from the IGME for their fruitful collaboration and discussion.

References

Ahuja, R. K., Magnanti, T. L. and Orlin, J. B., *Network Flows: Theory, Algorithms and Applications* (Englewood Cliffs, NJ: Prentice Hall, 1993).

Berardino, P., Fornaro, G., Lanari, R. and Sansosti, E. (2002), *A new algorithm for surface deformation monitoring based on small baseline differential interferograms*, IEEE Trans. on Geoscience and Remote Sensing *40*(11), 2375–2383.

Blanco, P., Mallorqui, J. J., Navarrete, D., Duque, S., Sanz-Marcos, J., Prats, P., Romero, R., Dominguez, J., Carrasco, D., Mora, O., Mallorqui, J. J. and Broquetas, T. (2005), *Application of the Coherent Pixels Technique to the Generation of Deformation Maps with ERS and ENVISAT Data*, Internat. Geosci. and Remote Sensing Symp., 2005. IGARSS '05., vol. 3, 1983–1986.

Blanco, P., Mallorqui, J. J., Duque S., and Navarrete, D. (2006), *Advances on DInSAR with ERS and ENVISAT data using the Coherent Pixels Technique (CPT)*, Proc. IGARSS 2006, Denver (USA).

Cima (2005), *Topographical leveling data of Lo Tacon industrial area in April 2003 and July 2004*, Dirección de Industria, Energia y Minas de la Consejera de Industria y Medio Ambiente de la region de Murcia.

Deffontaines, B., Fruneau, B. and Leparmentier, A. M. (2004), *Urban Instability Revealed by DINSAR and PS Interferometry: The Montmartre Case Example* (Paris, France), Proc. 2004 Envisat and ERS Symp. (ESA SP-572), Salzburg, (Austria). Published on CD-Rom.

European Space Agency (ESA) (2007), http://earth.esa.int/psic4/

Ferretti, A., Prati, C. and Rocca, F. (2000), *Nonlinear subsidence rate estimation using permanent scatters in differential interferometry*, IEEE Trans. *On Geoscience and Remote Sensing*, Vol. *38*, No. 5, pp. 2202–2212.

Ferretti, A., Prati, C. and Rocca, F. (2001), *Permanent scatterers in SAR interferometry*, IEEE Trans. *On Geoscience and Remote Sensing*, Vol. *39*, No. 1, pp. 8–20.

Ferretti, A., Savio, G., Barzaghi, R., Borghi, A., Sergio, M., Novali, F., Prati, C., and Rocca F. (2007), *Submillimiter accuracy of InSAR time series: Experimental validation*, IEEE Trans. *On Geoscience and Remote Sensing*, Vol. *45*, No. 5, pp. 1142–1153.

Fruneau, B., Carnec, C., Colesanti, C., Deffontaines, B., Ferretti, A., Mouelic, S., Le Parmentier, A. M., Rudant, J.P. (2003), *Conventional and PS differential SAR interferometry for monitoring vertical deformation due to water pumping: The Haussmann-St-Lazare case example* (Paris, France), Proc. FRINGE 2003, 1-5 December 2003, Frascati, (Italy).

Fruneau, B., Deffontaines, B., Rudanta, J. P., and Parmentier, A. M. (2005), *Monitoring vertical deformation due to water pumping in the city of Paris (France) with differential interferometry*, Comptes Rendus Geoscience *337*(13), September-October 2005, 1173–1183.

Gatelli, F., Guarnieri, A. M., Parizzi, F., Pasquali, P., Prati, C., and Rocca, F. (1994), *The wavenumber shift in SAR interferometry*, IEEE Trans. *Geosci. Remote Sensing*, Vol. 32, 855–865.

Hanssen, R., *Radar Interferometry. Data Interpretation and Error Analysis* (Kluwer Academic Publishers, 2001).

Herrera, G., Tomas, R., Lopez-Sanchez, J. M., Delgado, J., Mallorqui, J. J., Duque, S. and Mulas, J. (2007), *Advanced DInSAR analysis on mining areas: La Union case study (Murcia, SE Spain)*, Engin. Geol. Elsevier Sci. *90*, 148–159.

Herrera, G., Tomas, R., Lopez-Sanchez, J.M., Delgado, J., Vicente, F., Mulas, J., Altamira, Blanco, P., Duque, S., Mallorqui, J.J. (2008), *Comparison of advanced differential interferometry techniques: Murcia metropolitan area case study*, J. Photometry and Remote Sensing, submitted.

LANARI, R., MORA, O., MANUNTA, M., MALLORQUI, J.J., BERARDINO, P., and SANSOSTI, E. (2004), *A small-baseline approach for investigating deformations on full-resolution differential SAR interferograms*, IEEE Trans. *Geosci. Remote Sensing* 42, Issue 7, July 2004, pp. 1377–1386.

LANARI, R., CASU, F., MANZO, M., and LUNDGREN, P. (2007), *Application of the SBAS-DInSAR technique to fault creep: A case study of the Hayward fault, California*, Remote Sensing of Environment *109*(1), 12, 20–28.

LANARI, R., CASU, F., MANZO, M., ZENI, G., BERARDINO, P., MANUNTA, M., and PEPE, A. (2007), *An Overview of the small baseline subset algorithm: A DInSAR technique for surface deformation analysis*, Pure Appl. Geophys. *164*(4), 637–661.

LEE, D. T., and SCHACHTER, B.J. (1980), *Two algorithms for constructing a Delaunay Triangulation*, Int. J. Comp. Information Sci. *9*, 219–242.

MASSONET, D., ROSSI, M., CARMONA, C., ADRAGNA F., PELTZER, G., FEIGL, K., and RABAUTE, T. (1993), *The displacement field of the Landers earthquake mapped by radar interferometry*, Nature *364*, 138–142.

MONTI-GUARNIERI, A. and PRATI, C. (2000), *ERS-ENVISAT Combination for interferometry and super-resolution*, Proc. ERS-ENVISAT Symp. 2000.

MORA, O., MALLORQUI, J. J., and BROQUETAS, T. (2003), *Linear and nonlinear terrain deformation maps from a reduced set of interferometric SAR images*, IEEE Trans. Geosci. Remote Sensing *41*, 2243–2253.

MULAS, J. et al. (2003), *Geotechnical and hydrogeological analysis of land subsidence in Murcia (Spain)*, International Conf. Groundwater in Geological Engineering, Bled (Slovenia), Materials and Geoenvironment Groundwater in Geol. Eng. *50*, (1), 249–252.

NASA, The NASA/JPL SRTM web site. [Online]. Available: http://www.jpl.nasa.gov/srtm..

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W.T., and FLANNERY, B.P., *Numerical Recipes in c: the Art of Scientific Computing* (Cambridge University Press. 2002).

PELTZER, G. and ROSEN, P. A. (1995), *Surface displacement of the 17 Eureka valley, California, earthquake observed by SAR interferometry*, Science *268*, 1333–1336.

PIPIA, L., AGUASCA, A., FABREGAS, X., MALLORQUI, J. J., LOPEZ-MARTINEZ, C., and MARTURI, J. (2007), *Mining induced subsidence monitoring in urban areas with a ground-based SAR*, Proc. 2007 Urban Remote Sensing Joint Event, 11–13 April 2007, Paris.

SANZ-MARCOS, J., LOPEZ-DEKKER, P., MALLORQUI, J. J., AGUASCA, A. and PRATS, P. (2007), *SABRINA: A SAR bistatic receiver for interferometric applications*, IEEE Geosci. and Remote Sensing Lett. *4*(2), 307–311.

SARKAR, T. P. and RAO, S. M. (1984), *The application of the conjugate gradient method for the solution of electromagnetic scattering from arbitrarily oriented wire antennas*, IEEE Trans. Antennas and Propagation *32*, 398–403.

SIMONETTO, E., RAUCOULES, D., and CARNEC, C. (2005), *Interferometry of spatial radar images for post-mining surveillance: Experimental feedback*, Post Mining 2005 November 16-17, Nancy (France), CD ROM.

TOMAS, R., MARQUEZ, Y., LOPEZ-SANCHEZ, J. M., DELGADO, J., BLANCO, P., MALLORQUI, J. J., MARTINEZ, M., HERRERA, G., and MULAS J. (2005), *Mapping ground subsidence induced by aquifer overexplotation using advanced differential SAR interferometry: Vega media of the Segura River (SE Spain) case study*, Remote Sensing of Environement Elsevier Science *98*, 269–283.

(Received May 28, 2007, revised February 14, 2008, accepted February 14, 2008)

Pure appl. geophys. 165 (2008) 1195–1207
0033–4553/08/061195–13
DOI 10.1007/s00024-008-0353-5

Pure and Applied Geophysics

On the Occurence of Extreme Events in Long-term Correlated and Multifractal Data Sets

MIKHAIL I. BOGACHEV,[1] JAN F. EICHNER,[1] and ARMIN BUNDE[1]

Abstract—We review recent studies of the statistics of return intervals (i) in long-term correlated monofractal records and (ii) in multifractal records in the absence (or presence) of linear long-term correlations. We show that for the monofractal records which are long-term power-law correlated with exponent γ, the distribution density of the return intervals follows a stretched exponential with the same exponent γ and the return intervals are long-term correlated, again with the same exponent γ. For the multifractal record, significant differences in scaling behavior both in the distribuiton and correlation behavior of return intervals between large events of different magnitudes are demonstrated. In the absence of linear long-term correlations, the nonlinear correlations contribute strongly to the statistics of the return intervals such that the return intervals become long-term correlated even though the original data are linearly uncorrelated (i.e., the autocorrelation function vanishes). The distribution density of the return intervals is mainly described by a power law.

Key words: Return intervals, long-term correlations, multifractal records, nonlinear correlations, stretched exponential, power law.

1. Introduction

The understanding of the occurrence of extreme events is one of the major challenges in science. An important quantity here is the time interval between successive extreme events, and by understanding the statistics of these return intervals one aims to a better understanding of the occurrence of extreme events. In many cases, the extreme events do not appear to be independent, but instead they occur in clusters. Well-known examples in nature are temperature anomalies, extreme rainfalls and floods (PFISTER, 1998; GLASER, 2001; MUDELSEE *et al*., 2003; BUNDE *et al*., 2005). A clustering of extreme events has also been observed in processes related to human behavior, e.g., in stock prices or teletraffic in large networks (see BUNDE *et al*., 2002 and references therein).

Since the statistics of the occurrence times between extreme events in real systems is quite poor, one usually tries to extract information from events with rather smaller magnitudes that occur quite often and thus have enough statistics. The major issue is to find out some general "scaling" relations between the return intervals at low and high

[1] Institut für Theoretische Physik III, Justus-Liebig-Universität Giessen, 35392 Giessen, Germany.
E-mail: armin.bunde@uni-giessen.de

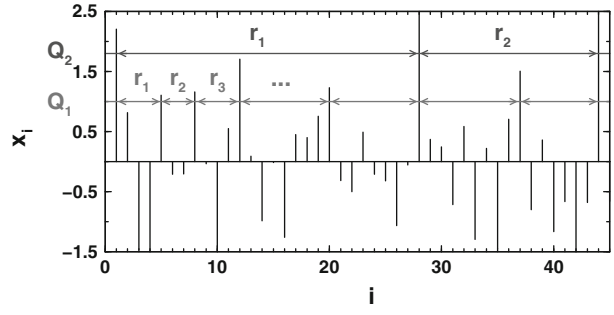


Figure 1
Illustration of the return intervals between events above two thersholds Q_1 and Q_2.

thresholds, which then allows extrapolation of the results to very large, extreme thresholds.

The object of our review is the return intervals r_j between single events x_i that exceed some fixed threshold Q (quantile). The process of retrieving a return interval series (r_j) with $j = 1, 2, ..., N_Q$ from a time series (x_i) of N data points following a distribution $P(x)$ is illustrated in Figure 1. Sometimes, instead of specifying the threshold Q, one specifies the mean return interval, or return period R_Q, since there is a one-by-one correspondence between both quantities, $R_Q = 1/\int_Q^\infty P(x)\,dx = N/N_Q$.

For a pure random process with statistically independent values with identical distribution, i.e., "i.i.d." data (e.g., Gaussian random white noise), also the return intervals are independent, and follow a Poisson distribution with a simple exponential probability density function $P_Q(r) = (1/R_Q)\exp(-r/R_Q)$ (see, e.g., v. Storch and Zwiers, 2001).

On the other hand, many processes in nature show long-term correlated behavior, either characterized by a single scaling exponent ("monofractal" behavior, see Bunde and Havlin, 1991; Koscielny-Bunde *et al.*, 1998; Eichner *et al.*, 2003; Kantelhardt *et al.*, 2003), or by a multitude of scaling exponents ("multifractal" behavior, see Bunde and Havlin, 1991; Mandelbrot, 1974; Lovejoy and Schertzer, 1991; Turcotte, 1992; Koscielny-Bunde *et al.*, 2006; Kantelhardt *et al.*, 2006). Long-term correlations strongly affect the statistics of extreme events, as was shown before in Bunde *et al.* (2005), and Eichner *et al.* (2006a), while the influence of multifractality has not been elaborated yet. In this work, we review former results for long-term correlated data sets (Bunde *et al.*, 2003, 2004, 2005) and our most recent findings for multifractal data sets (Bogachev *et al.*, 2007; Bogachev *et al.*, 2008; Bogachev and Bunde 2008a, b).

2. Return Intervals in Long-term Correlated Data Series

We consider a record (x_i), $i = 1, 2, ..., N$ and call the data long-term correlated, when the corresponding (linear) two-point autocorrelation function $C_x(s)$ decays by a power law,

$$C_x(s) = \frac{1}{\sigma_x^2 (N-s)} \sum_{i=1}^{N-s} (x_i - \langle x \rangle)(x_{i+s} - \langle x \rangle) \sim s^{-\gamma}, \quad (1)$$

where σ_x denotes the standard deviation, $\langle x \rangle$ the mean, and γ the correlation exponent, $0 < \gamma < 1$ of the data set. Such correlations are named "long-term" since the mean correlation time $T_\times = \int_0^\infty C_x(s)\, ds$ diverges in the limit of an infinitely long series (Bunde and Havlin, 1991). For uncorrelated x_i, $C_x(s) = 0$ for $s > 0$. If correlations exist up to a certain correlation time $s_\times$, then $C_x(s) > 0$ for $s < s_\times$ and $C_x(s) = 0$ for $s > s_\times$.

To generate long-term correlated data, we have used the Fourier-filtering technique, described, e.g., in Makse *et al.* (1996) and Schreiber and Schmitz, (1996). The power spectrum $P(f)$ of an uncorrelated random series with Gaussian distributed values is multiplied by $f^{-\beta}$. The series obtained by inverse Fourier transform of this modified power spectrum exhibits power-law correlations on all time scales.

Long-term correlated records are characterized by a distinct mountain-valley structure. Large values are rather followed by large values and small vaules are rather followed by small values. This feature of the correlated series (x_i) produces more numerous large intervals as well as substantially more small intervals, in comparison with uncorrelated records, changing the probability density function of the return intervals from an exponential to a "stretched" exponential, $\ln[P_Q(r)] \sim -(r/R_Q)^\gamma$, where the exponent γ is the correlation exponent (see Fig. 2) (Bunde *et al.*, 2003, 2004, 2005; Altmann and Kantz, 2005). For small values of r/R_Q the probability density function is described by a power-law, with an exponent close to $\gamma - 1$ (Eichner *et al.*, 2006b). When the data are shuffled, the correlations are destroyed, but the number N_Q of events above Q remains the same. Accordingly, for both correlated and uncorrelated records, R_Q is simply $R_Q = N/N_Q$, i.e., the return period R_Q is not affected by any correlations.

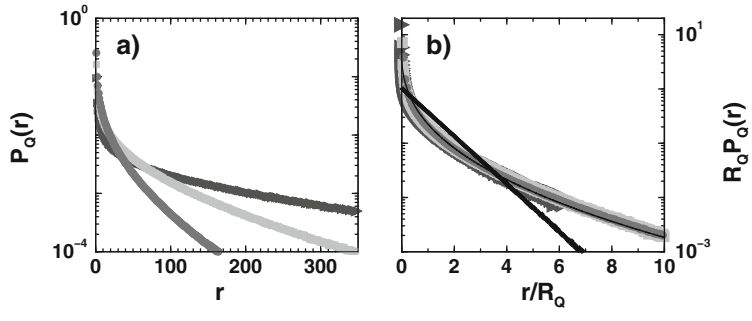


Figure 2

Distribution of the return intervals: (a) Displays the normalized distribution densities of the return intervals for three quantiles with $R_Q = 16$ (circles) 44 (squares) and 162 (triangles) (b) When scaling both axes appropriately, the curves collapse to a single curve, a stretched exponential of the form $R_Q P_Q(r) = b \exp[-a(r/R_Q)^\gamma]$, where γ is identical to the correlation exponent $\gamma = 0.4$ of the data. When shuffling the data, the curves follow a simple exponential (straight line in the semi-log plot) according to the Poisson statistics. Figure after … (Potsdam proc)

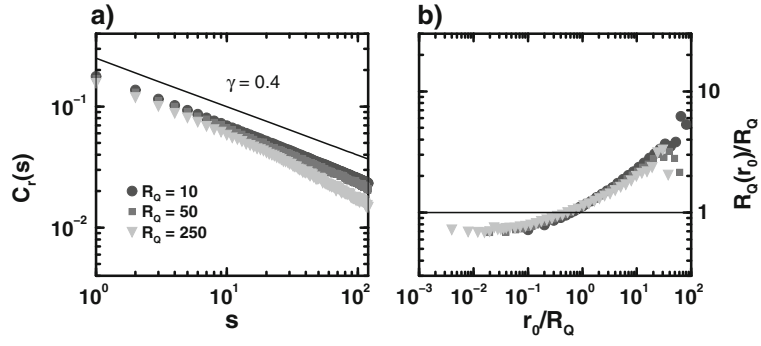


Figure 3

Long-term memory inherent in the return-interval series: (a) Displays the autocorrelation function $C_r(s)$ of the return intervals (r_j) achieved from long-term correlated Gaussian data with zero mean, unit variance, and $\gamma = 0.4$, for $R_Q = 10$, 50, and 250. The shape of the curves indicates a power-law behavior with the same exponent γ as the original data (the deviations for large R_Q and large scales s are due to finite-size effects.) A quantification of the resulting clustering effect is shown in figure (b). The conditional return period $R_Q(r_0)$, i.e., the mean return interval of all those r-values that immediately follow a return interval of a given size r_0, is plotted in units of R_Q versus r_0/R_Q. Figure after Bunde *et al.*, (2004).

The form of the distribution density $P_Q(r)$ reveals that very short and very long return intervals are more frequent than in uncorrelated data, but $P_Q(r)$ does not quantify if the return intervals themselves are arranged in a correlated fashion. In order to determine the autocorrelation behavior of the return interval series for a given quantile Q, we calculate the autocorrelation function $C_r(s)$ of the return interval series. Figure 3(a) shows $C_r(s)$ for data characterized by $\gamma = 0.4$ and three mean return periods $R_Q = 10$, 50, and 250. All curves exhibit a power-law decay $s^{-\gamma}$ with the similar slope, the correlation exponent γ, indicating long-term correlations corresponding to those present in the original data. Merely the curve for $R_Q = 250$ is slightly affected by finite-size effects. Indeed, these deviations clearly decrease with increasing data length N (not presented here). Due to the long-term correlations in the interval series, large return intervals are rather followed by large return intervals, and small return intervals are rather followed by small return intervals, leading on one hand to epochs with only slight extreme activity and on the other hand to a clustering of extreme events.

To quantify the effect of the long-range memory among the return intervals, we consider the conditional return intervals, i.e., we regard only those intervals whose preceding interval is of a fixed size r_0. In Figure 3(b) the conditional return period $R_Q(r_0)$, which is the mean of all conditional return intervals, is plotted versus r_0 in units of R_Q. The curves show a data collapse, similar as for the distribution densities, and display the memory inherent in the return interval series. Obviously, in the case of long-term correlations, the size of the expected return interval depends on the size of the preceding interval r_0. If the preceding interval r_0 is much larger than R_Q, the next return interval also will be much larger than R_Q, and *vice versa* for small intervals. This effect can cause huge

differences when predicting the occurrence of extreme events. If there is no memory in the data, i.e., no influence of the past r_0 on the upcoming return interval, then $R_Q(r_0)/R_Q \equiv 1$. All these effects can be seen in real data like climate data (measured, historical, and reconstructed) (BUNDE *et al*., 2004, 2005) or market volatility (YAMASAKI *et al*., 2005).

3. Generation of Multifractal Data Series

In multifractal data sets a single scaling exponent is not sufficient for a full description of the correlation structure of the data set, but rather an infinite number of exponents is needed (BUNDE and HAVLIN, 1991; FEDER, 1989). This happens, for example, when events of different magnitudes follow different scaling laws. To create data with such correlation structure, we employ a multiplicative cascade model, that allows us to create multifractal data with and without linear long-term correlations, depending on the parameters of the multipliers.

We consider a variant of the multiplicative random cascade process, described, e.g., in MANDELBROT *et al*., (1997). In the process, displayed in Figure 4, the data set is obtained in an iterative way, where the number of data points doubles in each iteration. We start with the zeroth iteration $n = 0$, where the data set $(x_i^{(0)})$ consist of one value, e.g., $x_1^{(0)} = 1$. In the n-th iteration, the data $x_i^{(n)}$, $i = 1, 2, \ldots, 2^n$, are obtained from the recurrent relation

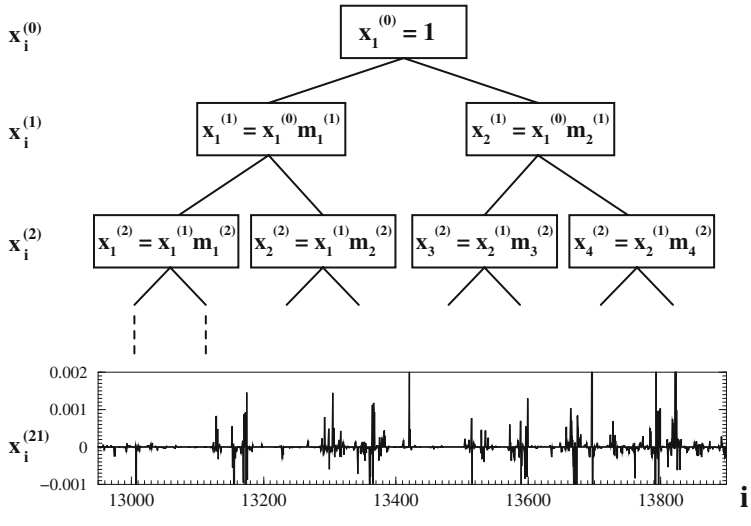


Figure 4

Illustration of the iterative random cascade process. After each iteration the length of the generated records is doubled and after $n = 21$ iterations the multifractal set consists of $N = 2^{21}$ numbers. An extract is shown in the bottom panel. Figure after BOGACHEV *et al*., 2008.

$$x_{2l-1}^{(n)} = x_l^{(n-1)} m_{2l-1}^{(n)} \quad \text{and} \quad x_{2l}^{(n)} = x_l^{(n-1)} m_{2l}^{(n)}, \tag{2}$$

where the multipliers $m_{2l-1}^{(n)}$ and $m_{2l}^{(n)}$ are random numbers following a log-normal distribution. Due to product stability the distribution of numbers, consisting of products of several random numbers, converges to a log-normal distribution, as for sum stability the distribution of the corresponding exponents converges to a Gaussian distribution. For this reason, we concentrate on log-normal multipliers in this work[1]. By altering the parameters $\langle m \rangle$ and σ_m, i.e., the mean and the standard deviation of the multipliers, the strength of the multifractality and the correlation properties of the data (x_i) are triggered.

There are several ways to characterize multifractal data sets. Here we chose the multifractal detrended fluctuation analysis (MF-DFA), introduced by KANTELHARDT *et al.* (2002). In the MF-DFA one considers the profile, i.e., the cumulated data series $Y_j = \sum_{i=1}^{j} (x_i - \langle x \rangle)$, and splits the record into N_s (non-overlapping) segments of size s. In each segment a local polynomial fit $y_v(j)$ of, e.g., second order is estimated. Then one determines the variance

$$F_v^2(s) = \frac{1}{s} \sum_{j=1}^{s} \left(Y_{[(v-1)s+j]} - y_v(j) \right)^2 \tag{3}$$

between the local trend and the profile in each segment v and determines a generalized fluctuation function $F_q(s)$,

$$F_q(s) \equiv \left\{ \frac{1}{N_s} \sum_{v=1}^{N_s} \left[F_v^2(s) \right]^{q/2} \right\}^{1/q}. \tag{4}$$

In general, $F_q(s)$ scales with s as $F_q(s) \sim s^{h(q)}$. For a monofractal time series, $h(q)$ is independent of q and identical to the Hurst exponent H (see, e. g., FEDER, 1989; HURST *et al.*, 1965)). For multifractal data, the generalized Hurst exponent $h(q)$ depends on the chosed moment q[2]. For $q = 2$, the relation $h(2) = 1 - \gamma/2$ holds for both, the monofractal and the multifractal case. In the absence of linear correlations (where $C_x(s) = 0$ for $s \geq 1$), $h(2) = 0.5$. In KANTELHARDT *et al.* (2002) it was shown, that $h(q)$ is directly related to the scaling exponent $\tau(q)$ defined by the standard partition function-based multifractal formalism (BUNDE and HAVLIN, 1991), via $\tau(q) = qh(q) - 1$.

To emphasize the presence or absence of linear correlations in the data series, we consider $F_q(s)/s^{0.5}$ (see Figs. 5(a), (c) and (e)), for $q = 0.5$, 1, 2, and 5, respectively. For multifractal data based on multipliers with $\langle m \rangle = 0.0$ (see Fig. 5(a)), the fitted slopes of the generalized fluctuation function are $h(q) - 0.5 = 0.67$, 0.34, 0.0, and -0.29,

[1] To create data with zero mean, we subtracted the mean of the generated log-normal record afterwards.

[2] The q-dependence of the Hurst exponent can be either due to the effect of the broad distribution of the data, or due to the nonlinear correlations inherent in the data, or due to both of the factors (KANTELHARDT *et al.* 2002). To ensure that the multifractality is created by nonlinear correlations and is not an artifact of a heavy-tail, MF-DFA can be applied after a rankwise exchange of the data with Gaussian distributed numbers (BOGACHEV *et al.*, 2007).

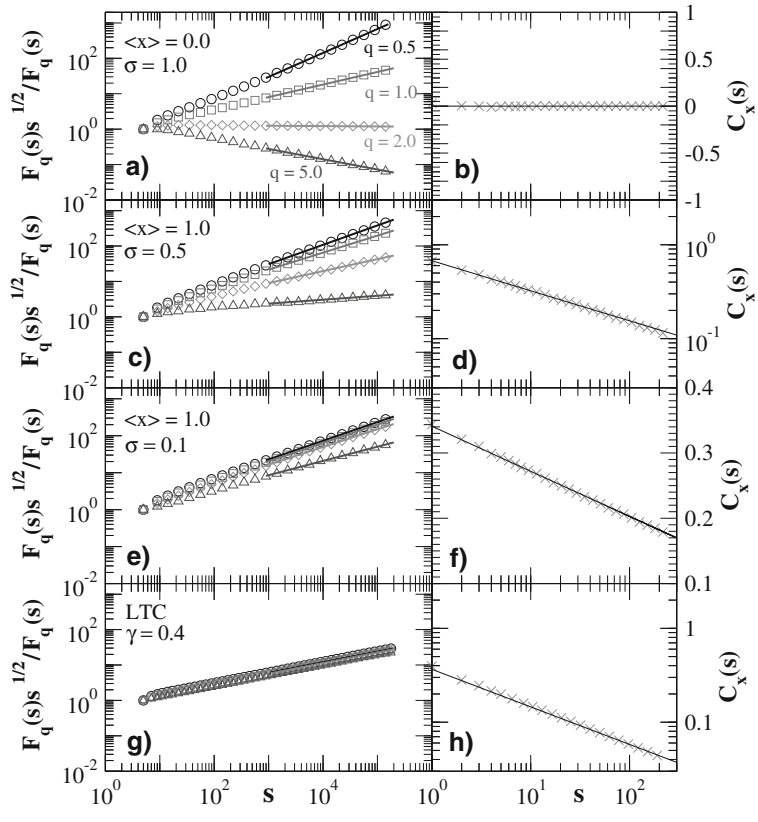


Figure 5

Detrended fluctuation analysis (MF-DFA) of the random cascade data. The left column contains results of the MF-DFA of data with $\langle m \rangle = 0.0$ and $\sigma_m = 1.0$ (a), $\langle m \rangle = 1.0$ and $\sigma_m = 0.5$ (c), $\langle m \rangle = 1.0$ and $\sigma_m = 0.1$ (e). The fitted slopes of the generalized fluctuation function $F_q(s)/s^{0.5}$ for the moments $q = 0.5$ (circles), 1 (squares), 2 (diamonds), and 5 (triangles) are shown. The right column presents autocorrelation functions for the three relevant cases, (b), (d) and (f), respectively. Note different scales for $C_x(s)$, depending on the kind of behavior we like to emphasize with getting nearly straight lines in either linear or logarithmic scale. For comparison, $F_q(s)/s^{0.5}$ and $C_x(s)$ for monofractal long-term correlated data with $\gamma = 0.4$ are given in (g) and (h), respectively.

respectively. Hence, this data set is strongly multifractal but apparently linearly uncorrelated (since $h(2) - 0.5 = 0$). To show that this result is not some kind of artifact of the MF-DFA method, we have confirmed it also by a direct calculation of the autocorrelation function $C_x(s)$, which is shown in Figure 5(b). In the case of data based on multipliers with $\langle m \rangle = 1.0$ and $\langle \sigma_m \rangle = 0.5$, the values for $h(q)$, shown in Figure 5(c), are 1.05, 0.98, 0.84, and 0.61, respectively. The corresponding autocorrelation function $C_x(s)$, shown in Figure 5(d), decays like a power-law with exponent $\gamma = 0.32$, corresponding to $h(2) = 0.84$ via $h(2) = 1 - \gamma/2$. In the case of data based on $\langle m \rangle = 1.0$ and $\langle \sigma_m \rangle = 0.1$, the multifractality is apparently weaker (the divergence of the slopes for different moments q is much less pronounced, see Fig. 5(e)), but $h(2)$ is close

to 1, attesting to almost non-stationary behavior. In this case, the autocorrelation function in Figure 5(f) decays logarithmically (straight line, fit: 0.34–0.03 ln(s)).

For comparison, we also show $F_q(s)/s^{0.5}$ and $C_x(s)$ for the Gaussian distributed monofractal data sets with $\gamma = 0.4$ in Figures 5 (g) and (h), respectively. The solid straight line in Figure 5 (h) represents the theoretical $C_x(s) \sim s^{-0.4}$. Conversely to the multifractal data sets, which are characterized by an infinite number of scaling exponents $h(q)$, the monofractal data sets are characterized by a single exponent. Hence, the generalized fluctuation functions $F_q(s)$ collapse after normalization, i.e., the exponents are independent of q, and correspond one-by-one to the exponent γ of the autocorrelation function by $h(q) \equiv h = 1 - \gamma/2$. We like to note that the differences between the values of the exponents of $F_q(s)$ are a necessary, but not a sufficient quantifier of the nonlinear correlations in the data, since they are also affected by the distribution of the data (KANTELHARDT *et al*. 2002). Therefore, straighforward application of the MF-DFA is not sufficient to distinguish between the distributional multifractality and nonlinear correlations, when dealing with non-Gaussian distributed data. Here we are not interested in the distributional part of multifractality, since it does not affect the statistics of return intervals (BOGACHEV *et al*., 2007). We claim that extracting return interval series from a data set is one of the ways to elucidate of the nonlinear long-term correlations inherent in the multifractal data.

4. Return Intervals in Multifractal Data Series

Next, we consider the return intervals. We begin with the probability density function $P_Q(r)$ of the return intervals. Figure 6 displays $R_Q\ P_Q(r)$ for three fixed return periods, $R_Q = 10$, 70, and 500. Contrary to the results based on pure long-term correlated (monofractal) data (see Fig. 2), the distribution densities for different R_Q follow neither a common curvature, nor a stretched exponential decay. For the multifractal data without linear correlations, where $\langle m \rangle = 0.0$ (see Fig. 6(a)), the distribution of the return intervals appears to decay by a clear power-law for all studied threshold values Q.

The fitted values of the exponents in Figure 6(a) are -1.35 for $R_Q = 500$, -1.59 for $R_Q = 70$, and -1.98 for $R_Q = 10$. For multifractal data with long-term correlations based on multipliers with $\langle m \rangle = 1.0$ and $\sigma_m = 0.5$ and $\sigma_m = 0.1$ (see Fig. 6(b) and (c)), there exist some deviations from the power-law behavior, which appear to increase with increasing linear long-term correlations (in Figure 6(c) the curvature of $R_Q\ P_Q(r)$ is more pronounced than in Figure 6(b), where the correlations are weaker). Rather, $R_Q\ P_Q(r)$ is some kind of superposition between a power-law and a stretched exponential, with different exponents for different quantiles Q. This shows that events exceeding quantiles of different sizes follow different statistical dependencies. After destroying the ordering of the original data by shuffling, the distribution densities for all R_Q-values display the expected exponential behavior (shown by the filled symbols in Fig. 6), as for shuffled monofractal data.

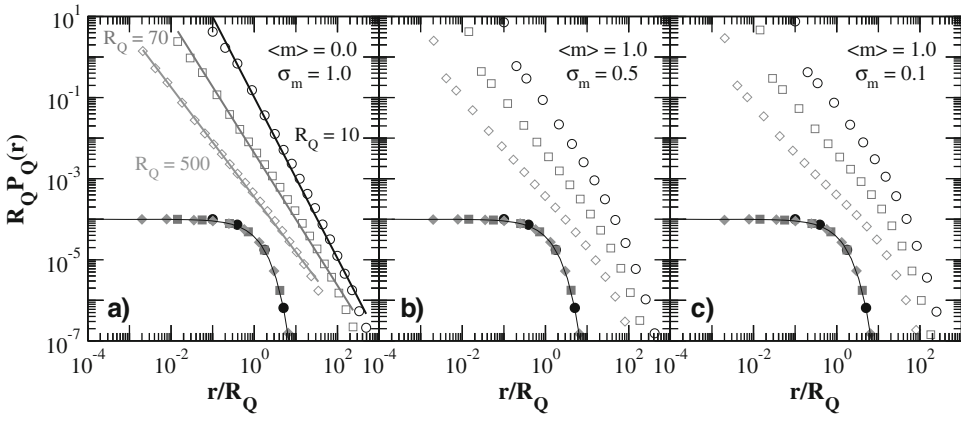


Figure 6

Scaled probability density function $P_Q(r)$ of the return intervals for the random cascade model based on multipliers with (a) $\langle m \rangle = 0.0$ and $\sigma_m = 1.0$, (b) $\langle m \rangle = 1.0$ and $\sigma_m = 0.5$, and (c) $\langle m \rangle = 1.0$ and $\sigma_m = 0.1$, for three different quantiles Q with return periods $R_Q = 10$ (open circles), 70 (open squares), and 500 (open diamonds). The curves for $R_Q = 70$ and 500 were shifted downwards by a factor of 10 and 100, respectively, to avoid overlapping symbols. When shuffling the data first, i.e., destroying the multifractality in the ordering of the data, the $P_Q(r)$ collapse to a single exponential curve (filled symbols, shifted downwards by a factor 10^5). All results were averaged over 150 configurations of original data sets of length $N = 2^{21}$.

We also studied the autocorrelation function $C_r(s)$ of the return interval series for multifractal data. In the absence of linear correlations, i.e., for data sets created with $\langle m \rangle = 0.0$, we find that $C_r(s)$ decays by a power-law (Fig. 7(a)), demonstrating the presence of long-term memory, even in the absence of linear correlations in the original data set. The exponents in the power-law exhibit a slight dependence on the size of the quantile Q, such that the intervals between smaller events (e.g., $R_Q = 10$) appear to be stronger correlated (i.e., show a smaller exponent γ) than the intervals between rather large events (e.g., $R_Q = 500$). This dependency on the quantile is a strong indicator for the influence of the multifractality in the data. For the case with inherent long-term correlations, i.e., data based on multipliers with $\langle m \rangle = 1.0$ and $\sigma_m = 0.5$, the return intervals themselves exhibit only slightly stronger long-term memory in comparison with the previous case (see Fig. 7(a)). It is not possible to characterize a scaling memory among the return intervals with a single exponent related to the γ of the autocorrelation function of the original data, as it holds for monofractal data. When changing the parameter σ_m to 0.1, the correlations in the original data approach the border of non-stationarity, but $C_r(s)$ displays only a weak effect on the values of the slopes (see caption of Fig. 7(c)). The effect of the multifractality on the slopes of $C_r(s)$ becomes slightly weaker, too, but remains dominant.

We also note, that the generation procedure we used is not the only way to create multifractal data. Another method, proposed in (Bacry et al. (2001), and usually referred

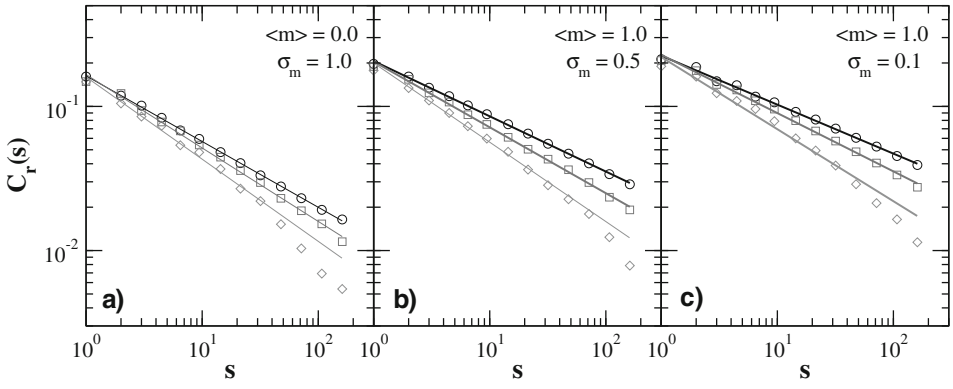


Figure 7

Autocorrelation analysis for the same data displayed in Figure 6, for the same three quantiles. Straight lines in the log-log plot indicate long-term correlations among the return intervals. In (a), the corresponding slopes are $-\gamma = -0.46$ (circles), -0.51 (squares), and -0.57 (diamonds), while in (b) they are -0.38, -0.46, and -0.55, and finally, in (c) they appear to be -0.34, -0.40, and -0.5, respectively. The deviations from the power-law behavior at large scales s for $R_Q = 500$ are due to finite-size effects. All results were averaged over 150 configurations of original data sets of length $N = 2^{21}$.

to as multifractal random walk (MRW), is also suitable to create multifractal data with vanishing linear ACF, but characterized by different values of the scaling exponents. We have also studied this model and obtained qualitatively similar results (Bogachev *et al.*, 2007; Bogachev and Bunde, 2008).

5. *Conclusion*

While for the monofractal long-term correlated data there is a direct relation between the correlation exponent γ of the data and both, distributional and time-ordering properties of the return intervals over a fixed threshold Q, for the multifractal data the situation is more complicated. For a multifractal data series without linear long-term correlations, the distribution of return intervals decays as a power-law with different exponents for different quantiles Q. When dealing with a data series that contains a superposition of multifractality and linear long-term correlations described by a correlation exponent $0 < \gamma < 1$, deviations from a power-law behavior are found, and they become stronger with increasing correlations, i.e., decreasing values of γ.

The most intriguing result, from our point of view, is the linear long-term correlated behavior of the return intervals for the multifractal data sets without any linear correlations. This reveals that both linear and nonlinear correlations present in the data set contribute to the linear correlations in the return intervals, such that even in the absence

of any linear correlations in the original data set the return intervals can be long-term correlated.

In this brief review we focused on calculations based on numerical models. We would like to mention that the results we described here also can be observed in real records. The clustering of the extreme events as well as the stretched exponential distribution that we described for long-term correlated records can be found, e.g., in annual temperatures of the Northern Hemisphere or the water levels of the river Nile (Bunde *et al*., 2004, 2005). Furthermore, the multifractal cascade model, on which this analysis was based, finds substantial applications in the simulation of returns in financial markets (Arneodo *et al*., 1998; Bouchaud *et al*., 2000; Fillol 2002; Muzy *et al*., 2006), and the results presented here can indeed be found in real economic data series (Bogachev *et al*., 2007; Bogachev and Bunde 2008a, b). We also wish to mention that a substantial application of our results can be found in the analysis of different physiological rhythms (Bogachev and Bunde 2008b), which demonstrate multifractal behavior (Stanley, 1999; Ivanov *et al*., 2001; Losa *et al*., 2005).

Acknowledgements

We benefited tremendously from discussions with our colleagues Shlomo Havlin and Jan W. Kantelhardt. Parts of their work on long-term correlated records have been reviewed in this article. We also extend appreciation for the partial support of the work which has been reviewed here from DAPHNet, and from the Deutscher Akademischer Austauschdienst (DAAD).

References

Altmann, E. G. and Kantz, H. (2005), *Recurrence time analysis, long-term correlations, and extreme events*, Phys. Rev. E *71*, doi:10.1103/PhysRevE.71.056106.

Arneodo, A., Muzy, J. F., and Sornette, D. (1998), *"Direct" causal cascade in the stock market*, Eur. Phys. J. B *2*, 277–282.

Bacry, E., Delour, J., and Muzy, J. F. (2001), *Multifractal random walk*, Phys. Rev. E *64*, doi:10.1103/PhysRevE.64.026103.

Bogachev, M. I., Eichner, J. F. and Bunde, A. (2007), *The effect of nonlinear correlations on the statistics of return intervals in multifractal data sets*, Phys. Rev. Lett. *99*, doi:10.1103/PhysRevLett.99.240601.

Bogachev, M. I., Eichner, J. F. and Bunde, A. (2008), *The effect of multifractality on the statistics of return intervals*, Eur. Phys. J. Special Topics *161*, 181–193.

Bogachev, M. I. and Bunde, A. (2008a), *Interoccurrence times between large returns in financial markets*, preprint.

Bogachev, M. I. and Bunde, A. (2008b), *Risk Estimations in Multifractal Records: Applications to physiology and Financing*, preprint.

Bouchaud, J.-P., Potters, M., and Meyer, M. (2000), *Apparent multifractality in financial time series*, Eur. Phys. J. B *13*, 595–599.

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*, doi:10.1103/PhysRevLett.94.048701.

BUNDE, A., EICHNER, J. F., KANTELHARDT, J. W., and HAVLIN, S. (2002), *The Science of Disasters – Climate Disruptions, Heart Attacks, and Market Crashes* (eds. Bunde, A., Kropp, J., and Schellnhuber, H.-J., Springer-Verlag, Berlin, 2002).

BUNDE, A. and HAVLIN, S. (1981), *Fractals and Disordered Systems*, (eds.: Bunde A., and Havlin, S., Springer-Verlag, Berlin, 1991).

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 of rare events in records with long-term persistence*, Physica A *342*, 308–314.

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*, doi:10.1103/PhysRevE.68.046133.

EICHNER, J. F., KANTELHARDT, J. W., BUNDE, A., and HAVLIN, S. (2006a), *Extreme value statistics in records with long-term persistence*, Phys. Rev. E *73*, doi:10.1103/PhysRevE.73.016130.

EICHNER, J. F., KANTELHARDT, J. W., BUNDE, A., and HAVLIN, S. (2006b), *Statistics of return intervals in long-term correlated records*, Phys. Rev. E *75*, doi:10.1103/PhysRevE.75.011128.

FEDER, J., *Fractals* (New York, Plenum, 1989).

FILLOL, J. (2003), *Multifractality: Theory and evidence an application to the French stock market*, Economics Bulletin *2*, No. 31, 1–12.

GLASER, R., *Klimageschichte Mitteleuropas* (Wissenschaftliche Buchgesellschaft, Darmstadt, 2001).

HURST, H. E., BLACK, R. P., and SIMAIKA, Y. M., *Long-term storage: An experimental study* (Constable & Co. Ltd., London, 1965).

IVANOV, P. Ch., AMARAL, L. A. N., GOLDBERGER, A. L., HAVLIN, S., ROSENBLUM, M. G., STANLEY, H. E., and STRUZIK, Z. R. (2001), *From 1/f noise to multifractal cascades in heartbeat dynamics*, Chaos *11*, 641–652.

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 behaviour*, 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. (Atmos.) *111*, D01106, doi:10.1029/2005JD005881.

KANTELHARDT, J. W., ZSCHIEGNER, S. A., KOSCIELNY-BUNDE, E., HAVLIN, S., BUNDE, A., and STANLEY, H. E. (2002), *Multifractal detrended fluctuation analysis of nonstationary time series*, Physica A *316*, 87–114.

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*, doi:10.1103/PhysRevLett.81.729.

KOSCIELNY-BUNDE, E., KANTELHARDT, J. W., BRAUN, P., BUNDE, A., and HAVLIN, S. (2006), *Long-term persistence and multifractality of river runoff records: Detrended fluctuation studies*, J. Hydrol. Vol. *322*, 120–137.

Fractals in Biology and Medicine (ed. by Losa, G. A., Merlini, D., Nannenmacher, T. F., and Weibel, E. R.) (Birkhäuser, Basel, 2005).

LOVEJOY, S. and SCHERTZER, D., *Scaling nonlinear variability in geodynamics: Multiple singularities, observals, universality classes*. In *Nonlinear Variability in Geophysics: Scaling and Fractals* (eds: Lovejoy, S., and Schertzer, D.) (Kluwer Academic Publ., Dordecht, Netherlands, 1991).

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*, doi:10.1103/PhysRevE.53.5445.

MANDELBROT, B. B. (1974), *Intermittent turbulence in self-similar cascades: Divergence of high moments and dimension of the carrier*, J. Fluid Mech. *62*, 331–358.

MANDELBROT, B. B., FISHER, A., and CALVET, L. (1997), *A multifractal model for asset returns*, Cowles Foundation Discussion Paper No. 1164, Working Paper, Yale University.

MUDELSEE, M., BÖRNGEN, M., TETZLAFF, G., and GRÜNWALD, U. (2003), *No upward trends in the occurrence of extreme floods in central Europe*, Nature, *425*, 166–169.

MUZY, J. F., BACRY, E., and KOZHEMYAK, A. (2006), *Extreme values and fat tails of multifractal fluctuations*, Phys. Rev. E *73*, doi:10.1103/PhysRevE.73.066114.

PFISTER, C., *Wetternachhersage, 500 Jahre Klimavariationen und Naturkatastrophen* 1496–1995, (Verlag Paul Haupt, Bern, 1998).

SCHREIBER, T., and SCHMITZ, A. (1996), *Improved surrogate data for nonlinearity tests*, Phys. Rev. Lett. *77*, doi:10.1103/PhysRevLett.77.635.

STANLEY, H. E., AMARAL, L. A. N., GOLDBERGER, A. L., HAVLIN, S., IVANOV, P. Ch. and PENG, C.-K. (1999), *Statistical physics and physiology: Monofractal and multifractal approaches*, Physica A *270*, 309–324.

TURCOTTE, D. L., *Fractals and Chaos in Geology and Geophysics* (Cambridge Univ. Press, 1992).

v. STORCH, H. and ZWIERS, F. W., *Statistical Analysis in Climate Research* (Cambridge Univ. Press, 2001).

YAMASAKI, K., MUCHNIK, L., HAVLIN, S., BUNDE, A., and Stanley, H. E. (2005), *Scaling and memory in volatility return intervals in financial markets*, PNAS *102*, 9424–9428.

(Received June 15, 2007, revised January 22, 2008, accepted February 6, 2008)

Pure appl. geophys. 165 (2008) 1209–1225
0033–4553/08/061209–17
DOI 10.1007/s00024-008-0354-4

Pure and Applied Geophysics

A Regional Archaeomagnetic Model for the Palaeointensity in Europe for the last 2000 Years and its Implications for Climatic Change

Fco. Javier Pavón–Carrasco,[1] Maria Luisa Osete,[1] J. Miquel Torta,[2] and Luis R. Gaya–Piqué[3]

Abstract—The SCHA.DI.00 directional model for the geomagnetic field in Europe for the last 2000 years (Pavón–Carrasco *et al*., 2008) has been updated by modelling the palaeointensity. This model, SCHA.DI.00, was developed from available Bayesian European Palaeosecular Variation Curves using the regional Spherical Cap Harmonic Analysis technique. The comparison of the palaeosecular variation curves, given by the regional model, with available archaeomagnetic data not used in its development showed an improvement with respect to the fit obtained by global archaeomagnetic models. In this paper advantage is taken of recently published palaeointensity databases to develop a complete (direction and intensity) regional archaeomagnetic model for the last 2000 years valid for the European region: the SCHA.DI.00–F model. Not only does this complete model provide an improvement for example for archaeomagnetic data studies, but it is also shown that this new regional model can be used to study the recently proposed link between the centennial secular variation of the geomagnetic field and climate change. The pattern of the archaeosecular variation of the field intensity obtained by SCHA.DI.00–F seems to verify the hypothesis presented by Gallet *et al.* (2005) about a possible (causal) connection between changes in the geomagnetic field intensity and in climate parameters, opening the door for more discussions on this challenging subject.

Key words: Palaeointensity, archaeomagnetism, geomagnetic secular variation, regional models, Europe.

1. Introduction

The long-term variation of the geomagnetic field extending over many years is called secular variation (SV). The temporal change of the geomagnetic field is far from linear, and abrupt changes in the rate of the secular variation change measured at the surface, known as geomagnetic jerks, provide information about the dynamics of the currents flowing in the Earth's outer core (Bloxham *et al*., 2002). When moving into the past, the Secular (or Palaeosecular) Variation Curves generated from archaeomagnetic

[1] Dpto de Geofísica y Meteorología, Universidad Complutense de Madrid, 28040, Madrid. E-mail: fjpavon@fis.ucm.es; mlosete@fis.ucm.es

[2] Observatori de l'Ebre, CSIC – Universitat Ramon Llull, Horta Alta 38, 43520. Roquetes, Spain. E-mail: jmtorta@obsebre.es.

[3] Equipe de Géomagnétisme, Institut de Physique du Globe de Paris, CNRS; Tour 14, 2 place Jussieu, 75005 Paris, France. E-mail: gaya@ipgp.jussieu.fr.

measurements can be used for dating purposes (e.g., LANOS, 2004). Moreover, a link between enhanced secular variation of the geomagnetic field and climate change over centennial time scales recently has been proposed (GALLET *et al.*, 2005, 2006; GALLET and GENEVEY, 2007; COURTILLOT *et al.*, 2007), challenging the role of solar forcing as the unique factor provoking these climatic variations.

The secular variation has been recorded directly through observatory measurements for the last two centuries. Declination (mostly) and inclination data are also available for the last four centuries from shipboard and navigational records (e.g., JACKSON *et al.*, 2000; JONKERS *et al.*, 2003). To extend our knowledge of the geomagnetic field variations into the past, palaeomagnetic studies are needed. The palaeosecular variation (PSV) in a region can be obtained from (a) heated archaeological structures, which are well-dated and not disturbed (archaeomagnetic curves), (b) well–dated volcanic materials, and (c) detailed sedimentary records (directional data or relative intensity). The use of archaeological material is normally preferred for several reasons: 1) The stability and origin of its remanence, commonly a thermo–remanence (TRM) or a partial thermo–remanence (pTRM); 2) the absence of delays in the remanence acquisition mechanism; 3) the stability of the carriers of the remanence; and 4) the facility of some archaeological materials to be accurately dated.

Archaeomagnetic data sets (e.g., the recent compilation of KORTE *et al.*, 2005) comprise directional and palaeointensity observations. The number of directional data is however higher (3787, about 7575 if we consider declination and inclination separately) than the amount of palaeointensity measurements (3206). In addition to this, archaeomagnetic data are not homogeneously distributed around the globe; Europe is the region where the highest record density is available.

Palaeomagnetic global models have been obtained during the last decade (e.g., OHNO and HAMANO, 1993; HONGRE *et al.*, 1998), and more recently by KORTE and CONSTABLE (2003, 2005) by using archaeomagnetic and sedimentary data. Since they are intended to represent the palaeofield on a global scale, these models are usually too smooth to record rapid changes of the Earth's magnetic field (i.e., archaeomagnetic jerks, GALLET *et al.*, 2005) which could be related to climatic changes. Recently the first directional regional model (SCHA.DI.00) to describe the palaeomagnetic field in Europe was proposed (PAVÓN–CARRASCO *et al.*, 2008), which seems to better reproduce the variability of the geomagnetic field over this region for the last 2000 years. The SCHA.DI.00 model was developed using the Spherical Cap Harmonic Analysis (SCHA) technique applied to five of the Bayesian European Palaeosecular Variation Curves (PSVC) (GALLET *et al.*, 2002; SCHNEPP and LANOS, 2005; MARTON and FERENCZ, 2006; GÓMEZ–PACCARD *et al.*, 2006a and ZANANIRI *et al.*, 2007), that are based on archaeomagnetic data. This model provided the directional behavior of the Earth's magnetic field, but no estimation about intensity was supplied because input data only contained directional information. The first spherical cap harmonic (SCH) coefficient, g_0^0, was used to normalize the rest of the coefficients.

In this study the present palaeointensity data set in Europe is used to adjust the first SCH coefficient, g_0^0, of the SCHA.DI.00 model to obtain a regional model for Europe

which also provides palaeointensity values for the last 2000 years. The new model, SCHA.DI.00–F, is compared to the existing global models of HONGRE *et al*. (1998), JACKSON *et al*. (GUFM, 2000), and KORTE and CONSTABLE (CALS7K.2, 2005) and with respect to real archaeointensity records. The use in the development of the regional model of a large amount of data over a restricted region of the planet makes it possible to achieve a higher spatial resolution compared to global models, and therefore the temporal variation of the field can be more accurately accounted for in that particular region. Finally, this new regional model can be used to study the recently proposed link between the centennial secular variation of the geomagnetic field and climate change (GALLET *et al*., 2005).

2. *The Previous SCHA.DI.00 Model*

The SCHA.DI.00 model (PAVÓN–CARRASCO *et al*., 2008) is based on the Spherical Cap Harmonic Analysis technique that was originally presented by HAINES (1985) and applied since to numerous geophysical studies (see Table 2 in TORTA *et al*., 2006, for a list of references). Although this method has been recently revised by THÉBAULT *et al*. (2006), the numerical problems are difficult to solve when only ground data are used, as in the case of archaeomagnetic data sets. Consequently the SCHA.DI.00 model used the classical approach of HAINES (1985), so the model cannot be extrapolated outside the limits of the cap. The SCHA algorithms of HAINES (1988) were adapted to the directional case by using Bauer's method (BARRACLOUGH, 1974), that relates the components X, Y and Z to the declination and inclination values. To produce a directional model (without intensity information) PAVÓN–CARRASCO *et al*. (2008) obtained a system of equations that depend on the declination (D) and inclination (I) data:

$$\begin{aligned}
&\sum_{k,m} G_{k,m}(\alpha_{k,m}\cdot \sin D - \beta_{k,m}\cdot \cos D) = 0\\
&\sum_{k,m} G_{k,m}(\alpha_{k,m}\cdot \sin I - \gamma_{k,m}\cdot \cos D\cos I) = \gamma_{0,0}\cos D\cos I\\
&\sum_{k,m} G_{k,m}(\beta_{k,m}\cdot \sin I - \gamma_{k,m}\cdot \sin D\cos I) = \gamma_{0,0}\sin D\cos I
\end{aligned} \tag{2.1}$$

where $\alpha_{k,m}, \beta_{k,m}, \gamma_{k,m}$ include the radial power, the colatitudinal Legendre, and longitudinal Fourier dependencies and $G_{k,m} = g_k^m/g_0^0$ are the SCH coefficients normalized to the first SCH coefficient g_0^0. This system of equations was introduced in the SCHA routines of HAINES (1988).

The input data of the SCHA.DI.00 model were the PSVC of Europe determined by application of hierarchical Bayesian modelling based on roughness penalty (LANOS, 2004). Five PSVC were used, which correspond to the regions of France, with its reference point located in Paris (GALLET *et al*., 2002); Germany, in Göttingen (SCHNEPP and LANOS, 2005); Hungary, in Budapest (MARTON and FERENCZ, 2006); Iberia, in Madrid (GÓMEZ–PACCARD *et al*., 2006a); and the United Kingdom, in Meriden (ZANANIRI *et al*., 2007). The databases

from Bulgaria (Kovacheva *et al.*, 1998 and references therein), Italy (Tema *et al.*, 2006; Korte *et al.*, 2005 and Tanguy *et al.*, 2003) and Austria (Schnepp and Lanos, 2006) were used to test the model.

The spherical cap expansion of SCHA.DI.00 extends up to $K_{int} = 2$ which, given the size of the spherical cap used (40° half angle), is equivalent, in terms of spatial wavelength, to a maximum degree of approximately 5 in the ordinary Spherical Harmonic Analysis (SHA).

3. SCHA.DI.00–F

The palaeointensity data used as input values for the development of the model were obtained from the global database of Korte *et al.* (2005). This database was updated with new data from Gallet *et al.* (2005) and Gómez–Paccard *et al.* (2006b) corresponding to French and Spanish locations. 611 palaeointensity determinations were available from different European countries for the time interval 0–1900 AD. The distribution of data in Europe is inhomogeneous, with high concentrations in France, England, Bulgaria, and Greece. Locations of the sites from which palaeointensity data have been used are shown in Figure 1.

The temporal distribution of data (Fig. 1, in box) shows a high density in the Roman period (between 100 and 300 AD). Similar to the archaeomagnetic directional database, there is a decrease in the density of data for 600–1000 AD (the so called "Dark Ages"). The palaeointensity data in Europe range between 36 and 106 μT; the average palaeointensity for 0–1900 AD is 60 μT. The mean palaeointensity error is 9 μT (maximum error of 21 μT and minimum of 1 μT). The mean time error is 42 years, with a minimum error of 0 and maximum of 300 years.

We have compared the palaeointensity data (F) compiled with the relative intensity data (f) provided by the directional regional model SCHA.DI.00 (Pavón–Carrasco *et al.*, 2008):

$$f^2 = x^2 + y^2 + z^2 = \left(\sum_{k,m} \alpha_{k,m} \cdot G_{k,m}\right)^2 + \left(\sum_{k,m} \beta_{k,m} \cdot G_{k,m}\right)^2 + \left(\sum_{k,m} \gamma_{k,m} \cdot G_{k,m}\right)^2 \tag{3.1}$$

where x, y and z are the relative Cartesian components of the geomagnetic field expansion in the SCH series, and $G_{k,m}$ represents the spherical cap harmonic coefficients g_k^m and h_k^m normalized to the first SCH coefficient, g_0^0. The relative intensity error can be obtained by applying the SCH coefficients errors to equation (3.1) (see Pavón–Carrasco *et al.*, 2008, Section 3).

From the relative intensity f calculated by the SCHA.DI.00 model at each location (Fig. 1) and for every epoch, the first SCH coefficient g_0^0 and its error were obtained as:

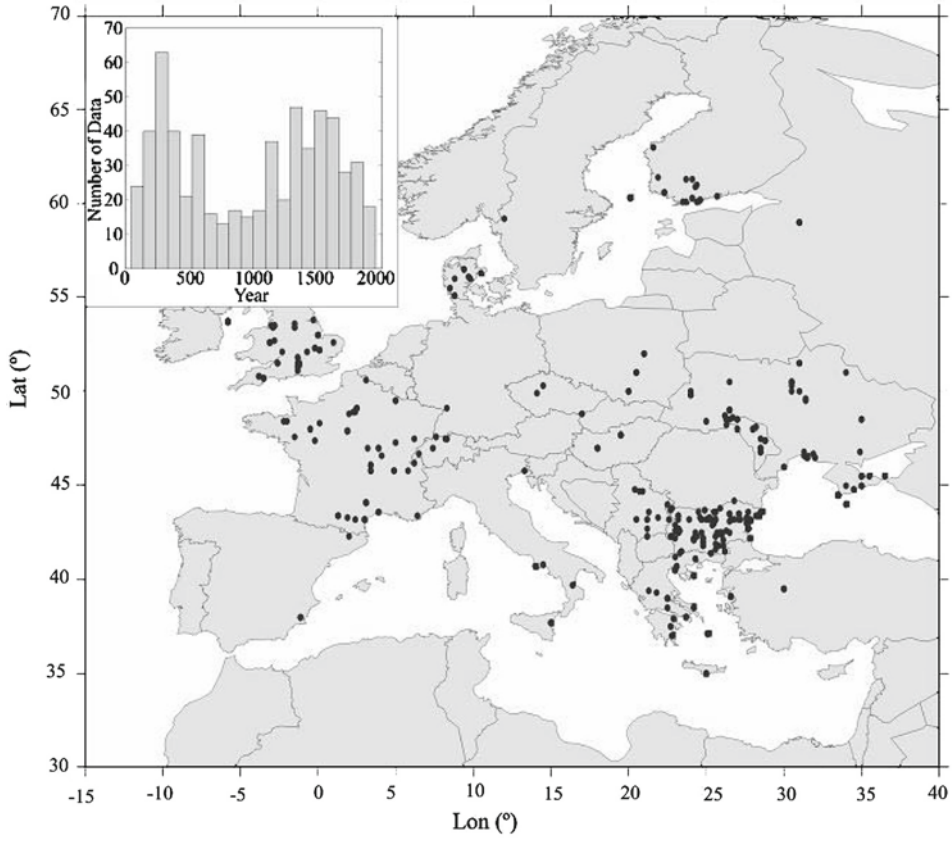


Figure 1
Map showing locations of palaeointensity data (KORTE *et al*., 2005; GALLET *et al*., 2005 and GÓMEZ-PACCARD *et al*., 2006b); and the temporal distribution of the data (in box).

$$g_0^0 = \frac{F}{f}; \quad \Delta g_0^0 = \frac{1}{f}\sqrt{F^2\Delta f^2 + f^2\Delta F^2}, \tag{3.2}$$

where F is the palaeointensity data and ΔF its error of the database of KORTE *et al*. (2005) and f is the relative intensity and Δf its error given by the SCHA.DI.00 model according to the equation (3.1). The g_0^0 coefficient has temporal but not spatial dependence, so these data sets were fitted using a time–dependent function. The temporal error of g_0^0 was considered to be equal to that of the palaeointensity data.

Considering that the mean time error of the palaeointensity data is 42 years, we have used windows of 50 years. The g_0^0 function was developed for each window by a polynomial temporal expansion (cubic splines) with knot points every 50 years. A weight function w_i was added in the inversion and is inversely proportional to the intensity and time errors:

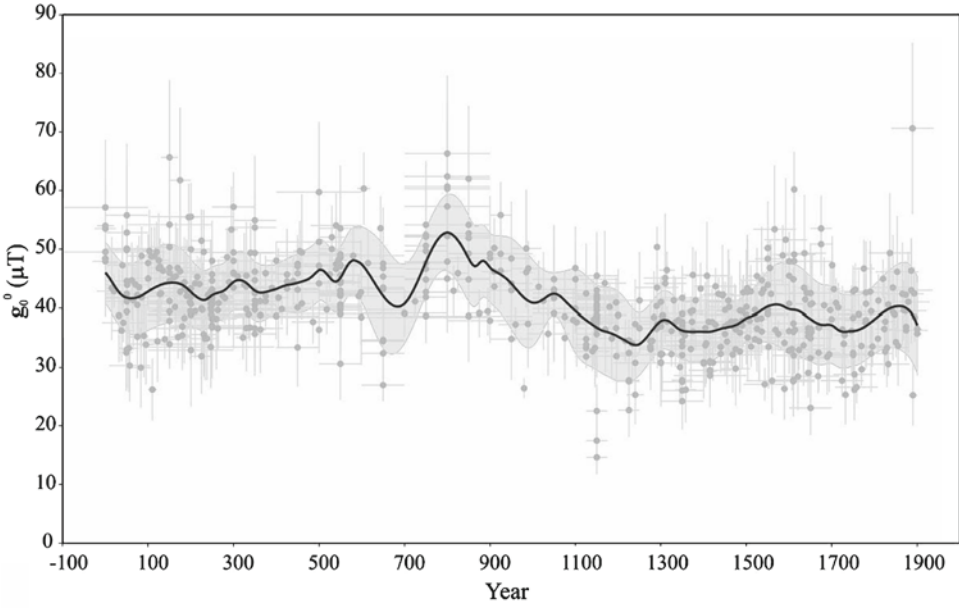


Figure 2
The first SCH coefficient, g_0^0. Dots: g_0^0 calculated by the global database (KORTE *et al.*, 2005; GALLET *et al.*, 2005 and GÓMEZ–PACCARD *et al.*, 2006b) with error bars. Curves: Fitted g_0^0 with error band.

$$w_i = \frac{1}{\rho_i + \tau_i}, \tag{3.3}$$

where ρ_i is the normalized error associated with the intensity and τ_i is the time normalized error. Once the g_0^0 function for the entire time interval is obtained, the rms (the square-root of the sum of the squared differences) was calculated by comparison with the calculated g_0^0 coefficient and input data.

Figure 2 shows the input data (the g_0^0 coefficient for each palaeointensity data measurement) and the g_0^0 time function and its error. The average value of the g_0^0 coefficient in Europe for the entire time interval is 41.5 μT, with a maximum at 800 AD. This age corresponds to abrupt changes in the magnetic field of the Earth, as pointed out by GALLET *et al.* (2005) and COURTILLOT *et al.* (2007).

Once the value of g_0^0 is obtained, we can derive the rest of SCH coefficients by using the expression $g_k^m = g_0^0 \cdot G_{k,m}$. Figure 3 shows the SCH coefficients and their associated errors. Coefficients g_0^0 and g_1^1 represent the main contribution to the Earth's magnetic field values. With this set of SCH coefficients it is possible to obtain the geomagnetic field components and intensity for Europe for the last 2000 years. The declination, inclination, and intensity values given by the SCHA.DI.00–F model for Europe every 100 years from 0 to 1900 AD are shown in the maps of Figure 4.

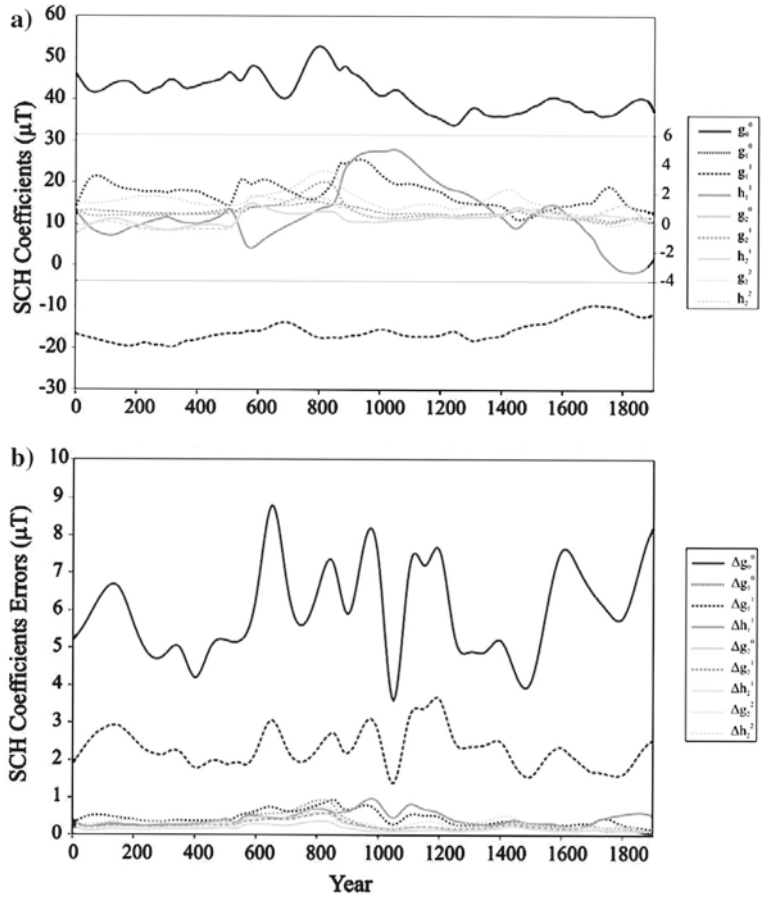


Figure 3
SCH coefficients (a) and their errors (b). The SCH coefficient errors are at 95% of confidence. Left scale corresponds to the SCH coefficients g_0^0 and g_1^1, right scale for the other SCH coefficients.

The new SCHA.DI.00–F and the previous directional SCHA.DI.00 models are available from the web site: http://pc213fis.fis.ucm.es/scha_model_f.html (palaeomagnetism group).

4. Discussion

4.1. The SCHA.DI.00–F Model

The palaeointensity values predicted by the SCHA.DI.00–F model have been compared with the *in situ* input data (we refer to Pavón–Carrasco *et al.*, 2008, for

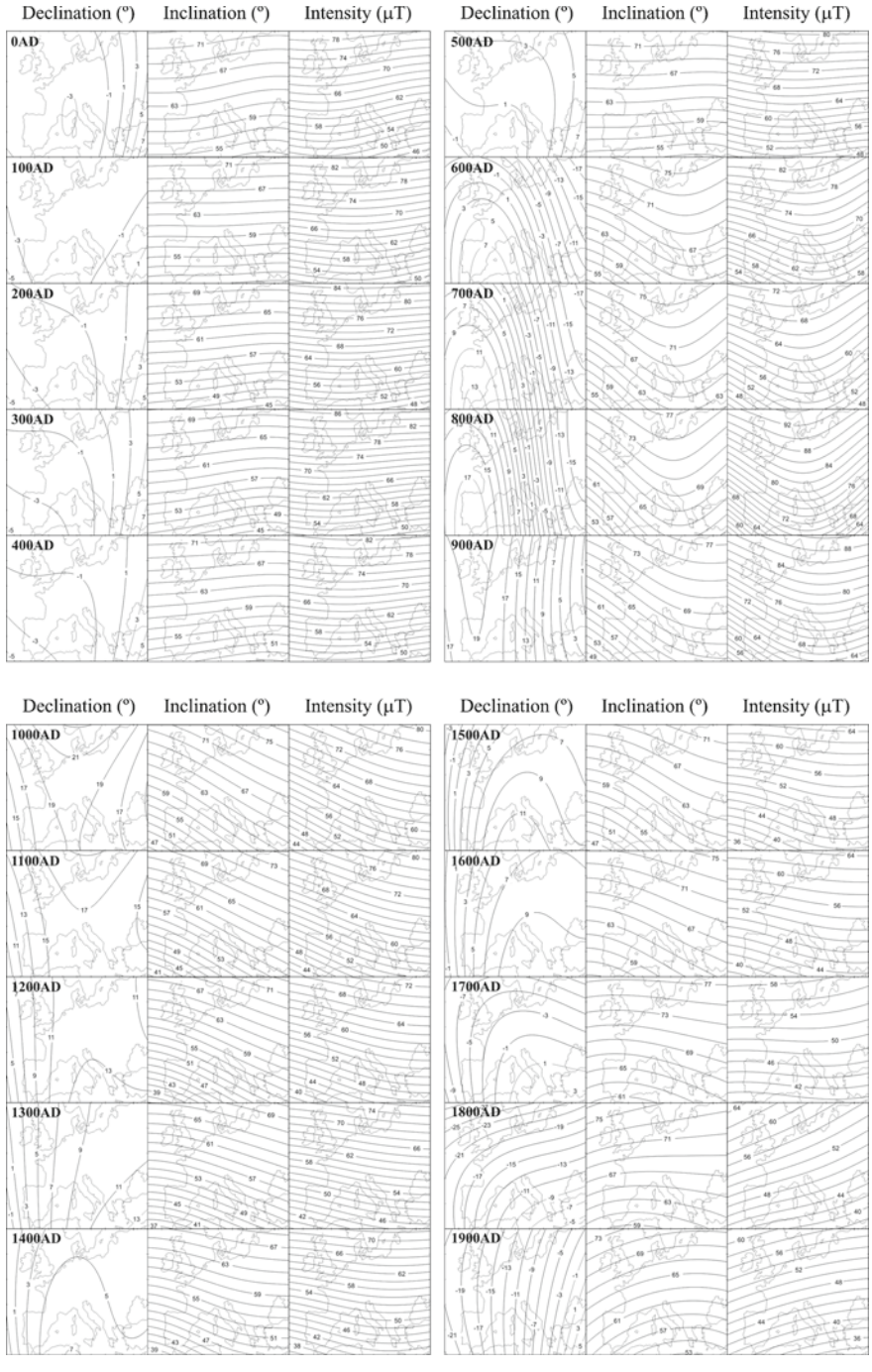


Figure 4
Declination, Inclination and Palaeointensity maps obtained by the SCHA.DI.00–F model at 100 year intervals.

directional data comparisons) and with the estimations from global models (CALS7K.2, KORTE and CONSTABLE, 2005–valid from 5000 BC to 1950 AD; GUFM, JACKSON *et al*., 2000–from 1590 to 1990 AD and HONGRE *et al*., 1998–from 0 to 1700 AD). The GUFM model only uses intensity data post 1840 AD and the authors assume a constant variation of the first SH coefficient ($\Delta g_1^0 = 15$ nT/year). Afterwards, GUBBINS *et al*. (2006) use the palaeointensity database of KORTE *et al*. (2005) from 1590 to 1840 AD to obtain the first SH coefficient g_1^0 in the same way that we have obtained the SCH coefficient g_0^0 in the present study. In this case, these authors propose a constant value in the variation of $g_1^0 = 2.28$ nT/year. For the comparison with the regional model, we have used the GUFM model with the new values of g_1^0 of GUBBINS *et al*. (2006) for the time period 1590–1840.

The error distribution for all these models has been plotted (Figs. 5a and 5b) with those obtained from SCHA.DI.00–F for comparison within the appropriate time period. For the period 0–1900 AD, the CALS7K.2 (KORTE and CONSTABLE, 2005) and the SCHA.DI.00–F models are compared. A total of 611 data have been used. The most frequent error is 5 μT for this regional model. In contrast, the error distribution of the global model exhibits a maximum at 7.5 μT. The mean quadratic error is very similar: 8.7 μT for the SCHA.DI.00–F model and 8.9 μT for CALS7K.2 model. The input data used in both models are also very similar. Therefore, the explanation of differences between the models resides in 1) the global model is also influenced by the data outside Europe and 2) the smoothing parameters used by KORTE and CONSTABLE (2005) seem to be too high to adequately describe brief, but significant Earth's magnetic field variations (this point is discussed later). The main differences between both models are the intervals 0–200, 800–950, and 1000–1100 AD, where the rms error of CALS model is higher than the regional model. Around 1300 and close to 1800 AD, the global model fits the data better than the regional model (Fig. 5b).

For the time interval 0–1700 AD, the distribution of errors of the SCHA.DI.00–F and the global model proposed by HONGRE *et al*. (1998) are shown in Figures 5a (center) and 5b. The mean quadratic errors are 8.7 μT for the regional model and 9.7 μT for the global model. In this case both distributions show a maximum in 5 μT, however the width of the error distribution is higher for the global model. It is important to consider that the input data for both models are very different in this case since many palaeointensity studies have been published in the last decade. This global model presents a poor fitting in the intervals 650–850 and 1400–1500 AD (Fig. 5b).

For the 1590–1900 AD interval the SCHA.DI.00–F model and the global GUFM (JACKSON *et al*., 2000) model are compared. For 1590–1840 AD we have used the modified GUFM model by GUBBINS *et al*. (2006). Mean quadratic error is 9.4 μT for the regional and 8.1 μT for the global model. The rms errors (Fig. 5b) show that the comportment of this global model is similar to the CALS global model for the considerate interval (1590–1900 AD), because the modified GUFM model also used the archaeointensity database of KORTE *et al*. (2005). In this case the global model seems to represent the behavior of the variation of intensity of the magnetic field better for this time period. However the GUFM model does not accurately describe the directional variation of the

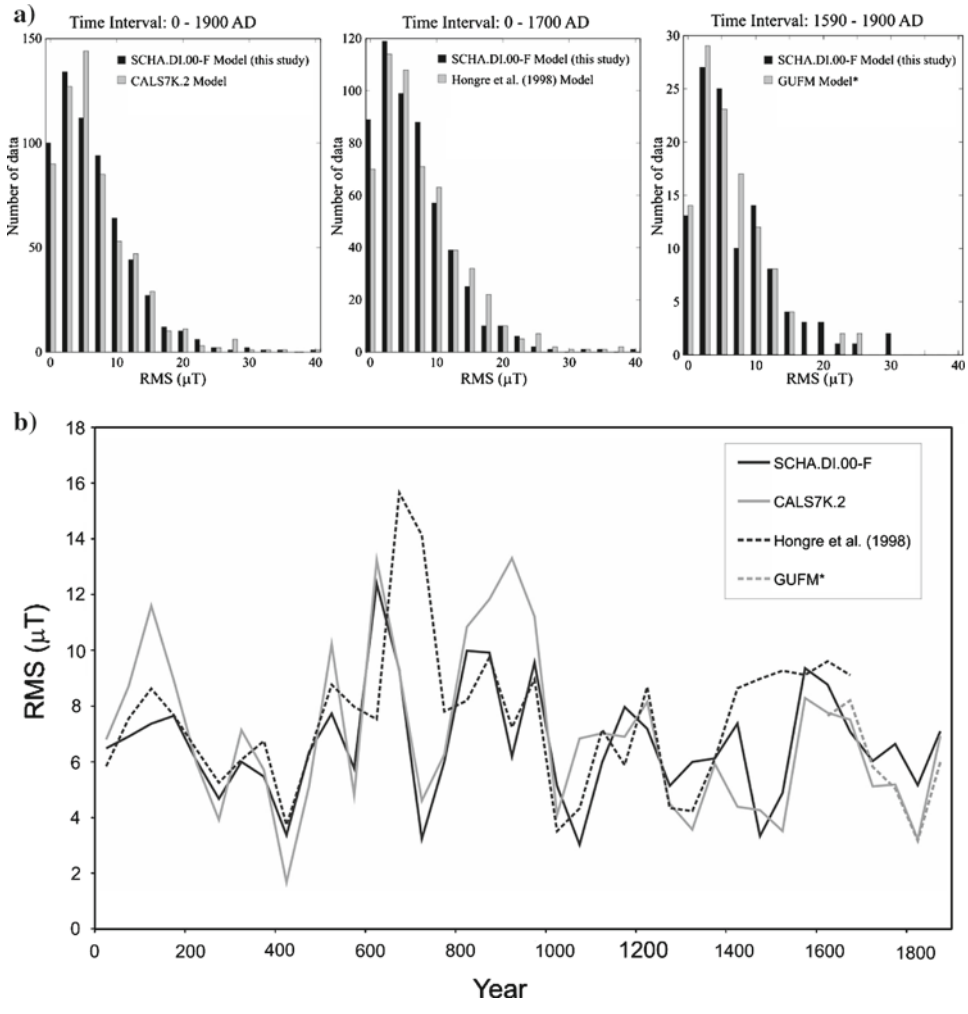


Figure 5

Histograms of errors (a) *versus* number of data and (b) *versus* time. Comparison of the rms errors between the SCHA.DI.00–F model and the global CALS7K.2 model (Korte and Constable, 2005); the global model of Hongre *et al*. (1998) and the global GUFM model (Jackson *et al*., 2000).* The GUFM model has been modified according to Gubbins *et al*. (2006).

geomagnetic field prior to 1700 AD (Pavón–Carrasco *et al*., 2008). Jackson *et al*. (2000) used historical directional observations of the magnetic field (from shipboards). The number of data used in the GUFM model was much higher (e.g., 83000 observations of magnetic declination before 1800 AD, Jackson *et al*., 2000) than those used in this study. For a detailed representation of the Earth's magnetic field a combination of both models should be considered in the future.

The SCHA.DI.00–F model suggests that the Earth's magnetic field in Europe reached 8 maximum peaks between 0–1900 AD at: 160, 320, 590, 820, 1070, 1310–1400, 1570 and 1770–1850 AD. Such a detailed description of the intensity variations of the geomagnetic field during the last 2000 years has not been achieved to date by any other geomagnetic model. This suggests that the smoothing parameters commonly used in constructing global models are too high to show such small wavelength variations.

4.2. Palaeointensity Generated Curves by SCHA.DI.00–F for France and Bulgaria

Palaeointensity data from Europe are inhomogeneously distributed (Fig. 1), most of the data being concentrated in France and Bulgaria. In Figure 6 the palaeointensity curves generated by the SCHA.DI.00–F model are compared with the palaeointensity data from France (KORTE *et al*., 2005 and references therein) after relocation to Paris, and from Bulgaria (KORTE *et al*., 2005 and references therein), relocated to Sofia (by the Virtual Axial Dipole Moment, VADM). The predicted palaeointensity curves generated by global models (KORTE and CONSTABLE, 2005; HONGRE *et al*., 1998 and JACKSON *et al*., 2000) are also shown.

The Bulgarian dataset suggests a higher variability in the intensity of the geomagnetic field than the French dataset. However, the French and Bulgarian datasets are only a part of the much larger amount of data used in the development of the models discussed in this paper. Consequently, they are influenced by the neighboring regions (as is the case of the regional model) or by data over the entire globe (global models). It also should be kept in mind that not all intensity values are determined using a unique method, therefore this could lead to variability in these datasets.

The CALS7K.2 global model (KORTE and CONSTABLE, 2005) is too smoothed to describe fluctuations shown by the data in these two locations. The model proposed by HONGRE *et al*. (1998) also seems to fail in isolating short–term fluctuations, whereas the GUFM model (JACKSON *et al*., 2000; GUBBINS *et al*., 2006) seems to represent the geomagnetic variations for the interval 1700–1900 AD most accurately.

The French and Bulgarian data appear to be in agreement with the SCHA.DI.00–F model except for two intervals. First, between 800–1000 AD, when a maximum in palaeointensity is predicted around 790–820 AD, which seems to be observed later in the Bulgarian database (about 900 AD). This location is mostly influenced by data from Ukraine (KORTE *et al*., 2005 and references therein). In addition, data from the Ukraine and Moldavia regions determined the position of the previous minimum at 690 AD. Second, a strong maximum is observed in the Bulgarian database around 1600–1650 AD (KOVACHEVA, 1997; KOVACHEVA *et al*., 1998), which is not well represented in the rest of the European database. The SCHA.DI.00–F shows a maximum at about 1570 AD but of lower magnitude.

Taking into account the European dataset used, it is suggested that future palaeointensity studies should be focussed in these two periods. The first one corresponds

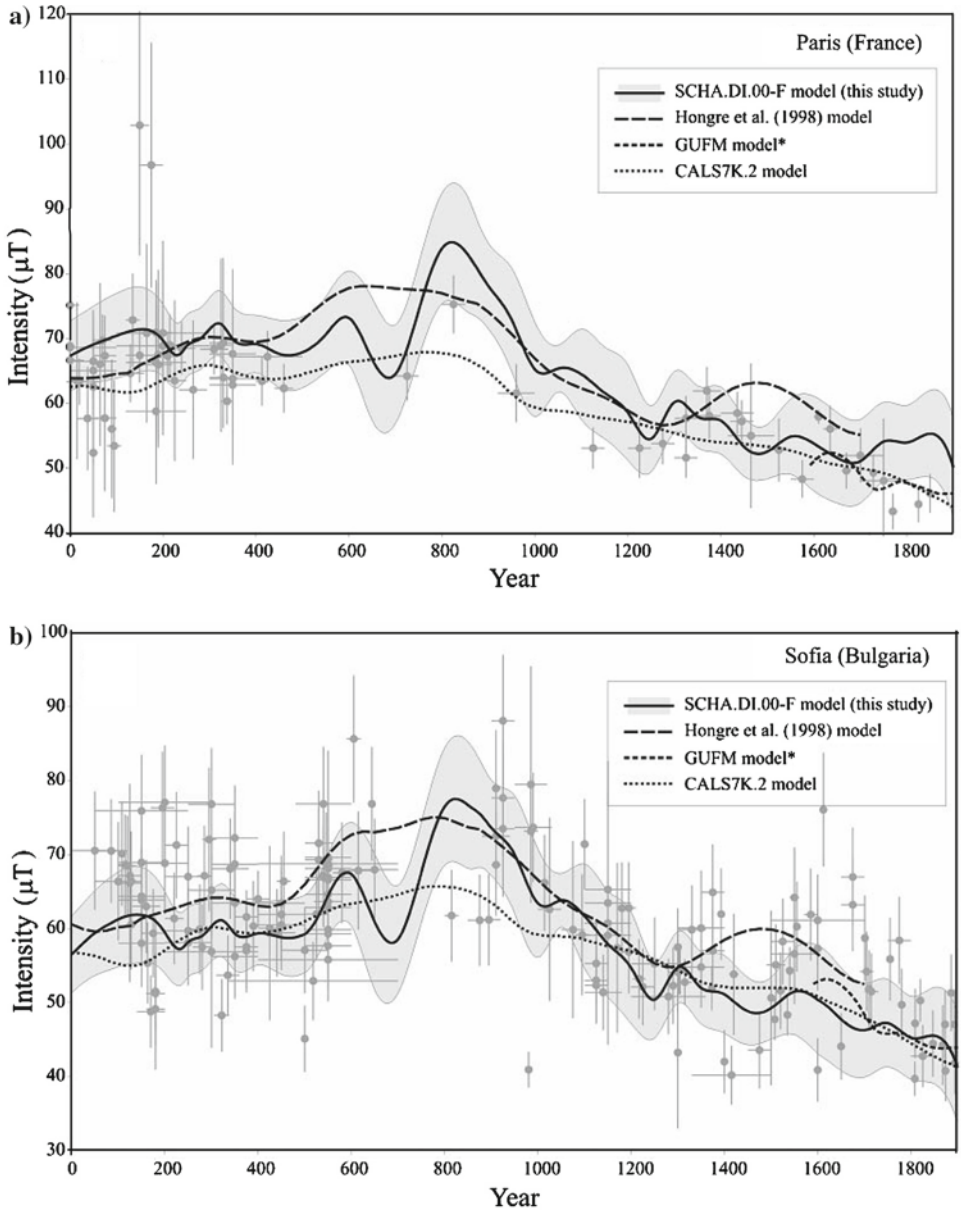


Figure 6

Palaeointensity curves for (a) Paris and (b) Sofia. Dots: Archaeointensity dataset from these regions with their error bars. Solid curve: Palaeointensity curve given by SCHA.DI.00–F model (with the error band) and the global geomagnetic models of Hongre *et al*. (1998), CALS7K.2 (Korte and Constable, 2005), and GUFM (Jackson *et al*., 2000).* The GUFM model has been modified according to Gubbins *et al*. (2006).

to the "Dark Ages" from which there is less palaeomagnetic information available (see Fig. 1, in box). The second (the maximum around 1600 AD observed in the Bulgarian database, KOVACHEVA, 1997; KOVACHEVA *et al*., 1998) seems to be poorly represented in Western Europe data.

4.3. Palaeointensity during the last 2000 Years and the Climatic Record

The last section of this paper deals with a hot topic of research — the relationship between the geomagnetic field and climate. Many studies have indicated that solar variability is one of the main non–anthropogenic sources for climate alterations in the past, because of the positive correlation between solar irradiance and the temperature at the Earth's surface (e.g., USOSKIN *et al*., 2005), at least until the decade of 1980 AD when anthropogenic causes are assumed to play an important role in climate change (e.g., LE MOUËL *et al*., 2005). The role of the Earth's magnetic field as an agent connected to climate variations has been discussed for decades, but recently the topic generated interest because more archaeomagnetic information are now available (see COURTILLOT *et al*., 2007 for a review). A plausible physical mechanism hypothesized to explain this connection is as follows: A change in the strength of the magnetic field would induce variability in the cosmic ray flux that reaches the troposphere, consequently modifying the rate of production of clouds and therefore altering the temperature at the Earth's surface (GALLET *et al*., 2006; COURTILLOT *et al*., 2007). However, many unanswered issues emerge when entering into details, such as the role of changes in the Earth's magnetic field compared either to CO_2 concentration or to the variation in the cosmic ray flux modulated by solar activity (COURTILLOT *et al*., 2007). GALLET *et al*. (2005) found a good agreement for Western Europe between cooling periods and archaeomagnetic jerks, defined as sharp increases in the intensity of the magnetic field contemporary to abrupt changes in its direction. The authors hypothesize that this may be a causal link, furthermore presenting these geomagnetic variations (a total of six possible archaeomagnetic jerks for the last two millennia) as the triggering events for climate variations which produced cultural changes in societies world wide (GALLET *et al*., 2006; GALLET and GENEVEY, 2007).

Figure 7 represents the palaeointensity curve for Paris generated by our model SCHA.DI.00–F, its error band, and the palaeointensity data from Western Europe (KORTE *et al*., 2005; GALLET *et al*., 2005 and GÓMMEZ–PACCARD *et al*., 2006b) relocated into the location of Paris by the VADM method. The errors associated with the palaeomagnetic measurements correspond to uncertainties in the intensity (vertical bar) and in the date (horizontal bar). The shaded stripes indicate cooling periods as deduced from the advance of the Swiss Alps glaciers (after HOLZHAUSER *et al*., 2005). Following the definition by GALLET *et al*. (2005), and taking into account the figures for the temporal evolution of the magnetic field direction presented in PAVÓN–CARRASCO *et al*. (2008), up to 8 archaeomagnetic jerks can be deduced from this curve. All of these coincide with a cooling period; in more detail, the rising part of each intensity

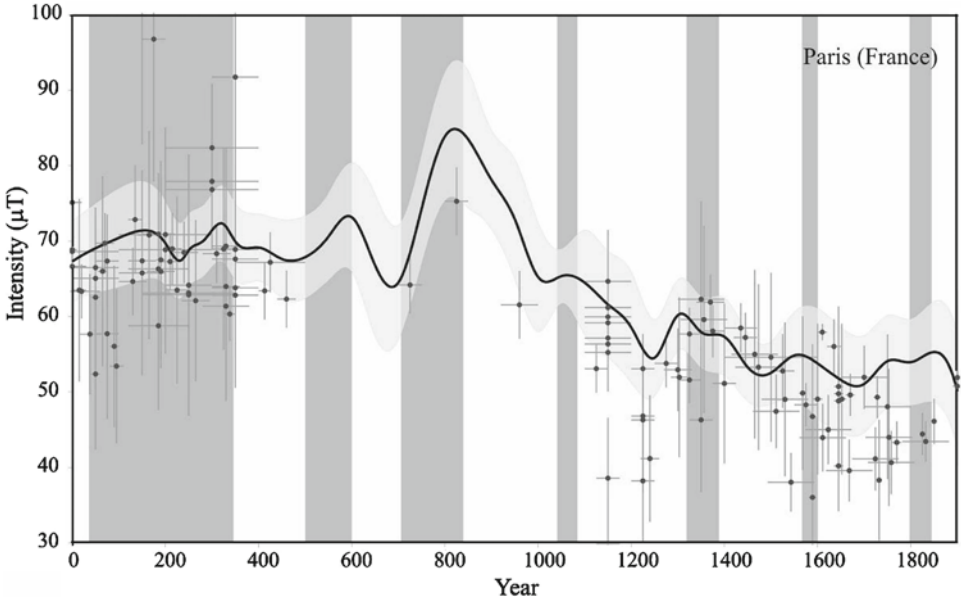


Figure 7
Palaeointensity curve at Paris (with error band) predicted by the SCHA.DI.00–F model. Archaeointensity data from Western Europe (KORTE *et al*., 2005; GALLET *et al*., 2005 and GÓMMEZ–PACCARD *et al*., 2006b) relocated to Paris by VADM method. Climatic variations summarized by GALLET *et al*. (2005) deduced from retreats and advances of the Alpine Glaciers for the past millennium studied by HOLZHAUSER *et al*. (2005). Cooling periods are indicated by shaded bands. Modified from GALLET *et al*. (2005).

maximum falls into a period of low temperature as shown by the shaded bands. Among these jerks, some of them (those around 820–950 AD, 1540–70 AD, and 1850 AD) are in good agreement with those found by GALLET *et al*. (2005), others (like the one for 1310–20 AD) precede the event detected by GALLET *et al*. (2005), but this event could extend up to 1400. The single jerk detected by GALLET *et al*. (2005) around 200 AD seems to split into two events as suggested by the new model, with maximum intensities and directional changes around 160–190 AD and 320 AD. The event around 590 AD suspected by GALLET *et al*. (2005) is better defined in the new model, therefore it is now considered to be a robust event. Finally, the low intensity maximum around 1070–1100 was not reported by GALLET *et al*. (2005) since no archaeointensity values were available from France for that epoch. The fact that more data are now being used helps to detect jerks more clearly, especially in the new model that incorporates data from other European regions.

If these eight events are real, the repetition time for the archaeomagnetic jerks has been about 250 years for the last 2000 years. This is usually assumed as the characteristic time for the non–dipolar part of the secular variation (e.g., HULOT and LE MOUËL, 1994; HONGRE *et al*., 1998). However, KORTE and CONSTABLE (2006) suggest a shorter–term

variability of the dipole field, thus this characteristic time sheds no light on either the dipolar or non–dipolar secular variation. The conclusion from this comparison is that there seems to be a correlation between geomagnetic field variation and climate, although a fuller cross correlation analysis is needed to test if this correlation is robust. This study allows no inference whether the link is causal or not. Neither can it be inferred that the archaeomagnetic jerks are of dipolar or non–dipolar origin. Such assessments require this type of study to be extended to other areas of the world, enabling other regional models to be developed to monitor the extent (global or regional) of this causal–noncausal relationship.

5. *Conclusions*

This paper shows how the inclusion of *in situ* palaeointensity data improves a regional archaeomagnetic model by providing a complete description of the geomagnetic field over a restricted area for the last 2000 years. Following a similar procedure, when substituting the PSVC directional input data by *in situ* directional data as well, the overall fit will be further improved in the future when a dense compilation will be finished.

The SCHA.DI.00–F model fits the present palaeointensity archaeomagnetic database for Europe more accurately than the global models proposed by HONGRE *et al*. (1998) and KORTE and CONSTABLE (2005) for the 0–1900 AD time interval. The regional model also fits the directional data properly (PAVÓN–CARRASCO *et al*., 2008), and, regarding the current data compilation, is the best model obtained to date over Europe for the 0–1900 AD time period. The model proposed by JACKSON *et al*. (2000) seems to represent most accurately the geomagnetic variations for 1700–1900 AD interval.

The new SCHA.DI.00–F model suggests that the Earth's magnetic field strength reached 8 maxima in Europe at: 160, 320, 590, 820, 1070, 1310–1400, 1570 and 1770–1850 AD. The complete model presented in this paper has also provided new insights into a very new and controversial topic of research, i.e., the question of whether connections exist between geomagnetic field changes and global (or regional) climate alterations. Although our results seem to confirm, even amplify, previous studies, it cannot be concluded that such relationships are statistically significant, nor can the existence of causality between both phenomena be considered established.

Acknowledgements

The authors are grateful to the Spanish research project CGL2005–00211, the FPI grant BES-2006-13488 and the IPGP contribution 2603. The paper benefited from the reviews of D. Tarling and A. Lodge, whose comments assisted the enhancement of the manuscript.

References

Barraclough, D.R. (1974), *Spherical Harmonic analyses of the geomagnetic field for eight epochs between 1600 and 1910*, Geophys. J. Int. *36*, 497–513.

Bloxham, J., Zatman, S., and Dumberry, M. (2002), *The origin of geomagnetic jerks,* Nature. *420*, 65–68.

Courtillot, V., Gallet, Y., Le Mouël, J–L., Fluteau, F., and Genevey, A. (2007), *Are there connections between the Earth's magnetic field and climate?* Earth Planet. Sci. Lett. *253*, 328–339.

Gallet, Y. and Genevey, A. (2007), *The Mayans: climate determinism or geomagnetic determinism?* EOS Trans. Am. Geophys. Un. *88*, N. 11, 129–130.

Gallet, Y., Genevey, A., and Le Goff, M. (2002), *Three millennia of directional variations of the Earth's magnetic field in western Europe as revealed by archaeological artefacts*, Phys. Earth Planet. Inter. *131*, 81–89.

Gallet, Y., Genevey, A., and Fluteau, F. (2005), *Does Earth's magnetic field secular variation control centennial climate change?*. Earth Planet. Sci. Lett. *236*, 339–347.

Gallet, Y., Genevey, A., Le Goff, M., Fluteau, F., and Eshraghi, S.A. (2006), *Possible impact of the Earh's magnetic field on the history of ancient civilizations*, Earth Planet. Sci. Lett. *266*, 17–26.

Gómez–Paccard, M., Lanos, Ph., Chauvin, A., McInstosh, G., Osete, M.L., Catanzariti, G, Ruiz–Martínez, V.C., and Núñez, J.I. (2006a), *The first archaeomagnetic secular variation curve for the Iberian Peninsula. Comparison with other data from Western Europe and with global geomantetic field models*, Geochemi., Geophys., Geosyst. *7*, Q12001, doi:10.1029/2006GC001476.

Gómez–Paccard, M., Chauvin, A., Lanos, Ph., Thiriot, J., and Jímenez–Castillo, P. (2006b), *Archeomagnetic study of seven contemporaneous kilns from Murcia* (*Spain*), Phys. Earth Planet Int. *157*, 16–32.

Gubbins, D., Jones, A.L., and Finlay, C.C. (2006), *Fall in Earth's magnetic field is erratic*. Science. *312*. 5775, 900–902.

Haines, G.V. (1985), *Spherical cap harmonic analysis*, J. Geophys. Res. *90* (B3), 2583–2591.

Haines, G.V. (1988), *Computer programs for spherical cap harmonic analysis of potential and general fields*, Comp. Geosci. *14*. 4, 413–447.

Holzhauser, H., Magny, M., and Zümbuhl, H. (2005), *Glacier and lake–level variations in west central Europe over the last 3500 years*, Holocene *15*, 789–801.

Hongre, L., Hulot, G., and Khokhlov, A. (1998), *An analysis of the geomagnetic field over the past 2000 years*, Phys. Earth Planet. Int. *106*, 311–335.

Hulot, G. and Le Mouël, J.L. (1994), *A statistical approach to the Earth's main magnetic field*, Phys. Earth Planet. Int. *82*, 167–183.

Jackson, A., Jonkers, A.R.T., and Walker, M.R. (2000), *Four centuries of geomagnetic secular variation from historical records*, Phil. Trans. R. Soc. Lond. A *358*, 957–990.

Jonkers, A.R.T., Jackson, A., and Murray, A. (2003), *Four centuries of geomagnetic data from historical records,* Rev. Geophys. *41*, 1006, doi:10.1029/2002R G000115.

Korte, M. and Constable, C. G. (2003), *Continuous global geomagnetic field models for the past 3000 years*, Phys. Earth Planet. Inter. *140*, 73–89.

Korte, M., Genevey, A., Constable, C.G., Frank, U., and Schnepp, E. (2005), *Continuous geomagnetic field models for the past 7 millennia: 1. A new global data compilation*, Geochem. Geophys. Geosyst. *6*, Q02H15, doi:10.1029/2004GC000800.

Korte, M. and Constable, C.G. (2005), *Continuous geomagnetic field models for the past 7 millenia: 2. CALS7K*, Geochem. Geophys. Geosyst. *6*, Q02H16, doi:10.1029/2004GC000801.

Korte, M. and Constable, C.G. (2006), *Centennial to millennial geomagnetic secular variation*, Geophys. J. Int. *167*, 43–52.

Kovacheva, M., (1997), *Archaeomagnetic database from Bulgaria: The last 8000 years*, Phys. Earth Planet. Int. *102*, 145–151.

Kovacheva, M., Jordanova, N., and Karloukovski, V. (1998), *Geomagnetic field variations as determined from Bulgaria archaeomagnetic data. Part II: The last 8000 years*, Sur. Geophys. *19*, 431–460.

Lanos, Ph., *Bayesian inference of calibration curves: Application to archaeomagnetism, in Tools for constructing chronologies: Crossing disciplinary boundaries*. (vol. 177, eds. C. Buck, and A. Millard, 2004) pp. 43–82 (Springer–Verlag, London 2004).

LE MOËL, J.–L., KOSSOBOKOV, V., and COURTILLOT, V. (2005), *On long–term variations of simple geomagnetic indices and slow changes in magnetospheric currents; the emergence of anthropogenic global warming after 1990?* Earth Planet. Sci. Lett. *232*, 273–286.

MARTON, P. and FERENCZ, E. (2006), *Hierarchical versus stratification statistical analysis of archaeomagnetic directions: The secular variation curve for Hungary*. Geophys. J. Int. *164*, 484–489.

OHNO, M. and HAMANO, Y. (1993), *Spherical harmonic analysis of palaeomagnetic secular variation curves*, Central Core Earth *3*, 205–212.

PAVÓN–CARRASCO, F.J., OSETE, M.L., TORTA, J.M., GAYA–PIQUÉ, L.R., and LANOS, Ph. (2008), *Initial SCHA.DI.00 regional archaeomagnetic model for Europe for the last 2000 years*, Phys. Chem. Earth A/B/C *33*, 6–7, 596–608.

SCHNEPP, E. and LANOS, Ph. (2005), *Archaeomagnetic secular variation in Germany during the past 2500 years*, Geophys. J. Int. *163*, 479–490.

SCHNEPP, E. and LANOS, Ph. (2006), *A preliminary secular variation reference curve for archaeomagnetic dating in Austria*, Geophys. J. Int. *166* (1), 91–96.

TANGUY, J. C., LE GOFF, M., PRINCIPE, C., ARRIGHI, S., CHILLEMI, V., LADELFA, S., and PATANE, G. (2003), *Archeomagnetic dating of Mediterranean volcanics of the last 2100 years: Validity and limits*, Earth Planetary Sci. Lett. *211*, 111–124.

TEMA, E., HEDLEY, I., and LANOS, Ph. (2006), *Archaeomagnetism in Italy: A compilation of data including new results and a preliminary Italian secular variation curve*, Geophys. J. Int. *167*, 1160–1171.

THÉBAULT, E., SCHOTT, J.J., and MANDEA, M. (2006), *Revised spherical cap harmonic analysis (R–SCHA): Validation and properties*, J. Geophys. Res. *111*, B01102, doi: 10.1029/2005JB003836.

TORTA, J.M., GAYA–PIQUÉ, L.R., and DE SANTIS, A. (2006), *Spherical cap harmonic analysis of the geomagnetic field with application for aeronautical mapping*. In Rasson, J.L. and Delipetrov, T., eds., *Geomagnetics for Aeronautical Safety: A Case Study in and around the Balkans*, NATO Security Through Science Series–C, 291–307.

USOSKIN, I.G., SCHÜSSLER, M., SOLANKI, S.K., and MURSULA, K. (2005), *Solar activity, cosmic rays, and Earth's temperature: A millennium–scale comparison*, J. Geophys. Res. *110*, A10102, doi:10.1029/2004JA010946.

ZANANIRI, I., BATT, C.M., LANOS, Ph., TARLING, D.H., and LINFORD, P. (2007), *Archaeomagnetic secular variation in the UK during the past 4000 years and its application to archaeomagnetic dating*, Phys. Earth Planet. Inter. *160*, 2, 97–107.

(Received June 9, 2007, revised January 24, 2008, accepted January 24, 2008)